Genome skimming for phylogenomics

Steven Andrew Dodsworth

School of Biological and Chemical Sciences,
Queen Mary University of London,
Mile End Road,
London E1 4NS, UK.

Submitted in partial fulfilment of the requirements of the degree of

Doctor of Philosophy

November 2015

Statement of originality

I, Steven Andrew Dodsworth, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged and my contribution indicated. Previously published material is also acknowledged and a full list of publications is given in the Appendix. Details of collaboration and publications are given at the start of each chapter, as appropriate.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: 

Date: 16th November 2015

Frontispiece: *Nicotiana burbidgeae* Symon at Dalhousie Springs, South Australia. 2014. Photo: S. Dodsworth.

Acknowledgements

Firstly, I would like to thank my PhD supervisors, Professor Andrew Leitch and Professor Mark Chase. I thank Andrew for his continual daily encouragement and infectious enthusiasm; I am eternally grateful to have had him as a principal supervisor and mentor for the past three years. I thank Mark for a true sense of botanical inspiration, and particularly for his excitement and ideas regarding *Nicotiana* section *Suaveolentes*.

I acknowledge NERC, the Linnean Society, the Systematics Association and Botanical Research Fund for support I have received. Various colleagues in London and further afield have been important in shaping my ideas: Xavier Aubriot, Bill Baker, Jurriaan de Vos, Laura Kelly, Ilia Leitch, Jamie McCann, Jiří Macas, Alex Monro, Petr Novák, Stephen Rossiter, Tiina Särkinen, Hanna Schneeweiss.

In our lab at QM I would like to sincerely thank my partner in crime Maïté Guignard, along with Hannes Becher, Wencai Wang, and many other members of the 5th floor labs. I owe huge thanks to Monika Struebig for her daily assistance in the lab, without whom huge chunks of this thesis would not have been feasible. For provocative discussions on Solanaceae, and access to BM specimens, I sincerely thank Sandy Knapp.

Phenomenal fieldwork assistance and an effervescence regarding *Nicotiana* was provided by Maarten Christenhusz, John Conran and Wayne England. Friends and family have spurred me on, in particular Rachel Walker and Juliana Oladuti. Finally, I am indebted to my partner Gbemiga Oladuti for putting up with the incessant strains and stresses that a doctoral thesis inevitably entails, and for helping me to focus on the light at the end of the tunnel.

Abstract

The advent of next-generation (or high-throughput) sequencing (NGS/HTS) has revolutionised biology, with much impact on the field of molecular phylogenetics. Traditional debates of taxa versus characters are now somewhat defunct in the phylogenomics era. In this thesis I focus on one particular HTS approach, 'genome skimming' as a phylogenomics and genomics method. I extend the scope of genome skimming to encompass more of the data present from low-coverage genome sequencing, using a novel method to analyse genomic repeat abundances as phylogenetic characters in addition to the assembly of high-copy organellar and nuclear DNA (plastomes and the nuclear ribosomal DNA cistron). The methodology for using nuclear repeats is initially developed, and then genome skimming is used to explore the phylogenetic relationships within a recent radiation – *Nicotiana* section *Suaveolentes* (Solanaceae). These data provide a significant improvement in our phylogenetic understanding of the group, despite low levels of genetic divergence between the core Australian species of *Nicotiana* section *Suaveolentes* and significant incomplete lineage sorting. Support is garnered for the whole genome duplication (WGD) radiation lag-time model in section *Suaveolentes*, with a significant increase in diversification in the last 2 million years following a lag of approximately 4 million years after the origin of the section at ~6.8 mya (allopolyploidisation event). Associated with this diversification are various processes of diploidisation including chromosome number reduction and genome downsizing. In addition to genomic patterns, there are ecological ones associated with diversification, including a general switch from perennial to annual life history strategy (with some notable reversals). These results paint *Nicotiana* section *Suaveolentes* as a recent and ongoing radiation, and are placed in the broad context of angiosperm diversification post-polyploidisation.

Contents

List of Figures

List of Tables

Chapter 1 General Introduction

Publication information

Some of the ideas presented in this chapter are published in the following article, for which I was sole author.

Dodsworth S. (2015). Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science*, 20: 525-27.

Molecular phylogenetics and molecular markers

Key advances in the late 1980s and early 1990s led to a surge in the development of molecular phylogenetics methods and data in modern systematics. The advent of the polymerase chain reaction (PCR) coupled with advances in Sanger sequencing made it possible to sequence coding and non-coding regions of genomic DNA (gDNA) with relative ease (Mullis *et al.*, 1986; Smith *et al.*, 1986; Saiki *et al.*, 1988); the former enabled specific amplification of genes and regions of interest and the latter, in particular once automated detection of fluorophores became the norm, enabled sequencing of up to ~800 bp of DNA in each reaction. These generic advances in molecular biology coupled with ones more specific to plants, e.g. the use of silica gel to preserve material for later DNA studies (Chase and Hills, 1991), further widened the scope of molecular phylogenetics to include many and rare taxa from across the globe.

Due to the ease of amplification with relatively little modification to PCR protocols, the focus of much early work was DNA present in high-copy number – either in the genome or cell. Specifically, this included the nuclear internal transcribed spacer of ribosomal DNA (ITS) and organellar DNA (mitochondrial and plastid genomes). In animals this quickly led to a surge in the use of cytochrome oxidase subunit I (*cox1*) from the mitochondrial genome (mitogenome) as the marker of choice for systematic studies (e.g. Hunt *et al.*, 2007). In plants, mitogenomes are much larger (typically 400 kb or larger, compared to 15-18 kb in animals), evolve slower, and contain many and more frequent genome rearrangements (Palmer 1992). These facets of plant mitogenomes together made their use in plant systematics less feasible on the whole, although they may have use in deeper level systematics (Turmel *et al.*, 2002; Knoop 2004). Instead the plastid genome (plastome) became the focus of plant molecular systematics. By contrast the plastome is typically around 150 kb, harbouring fewer rearrangements, and yet approximately a four-fold increase in the rate of nucleotide substitution compared to the mitogenome. The

coding genes *rbcL* (ribulose bisphosphate carboxylase large subunit) and *matK* (maturase K) became the markers of choice, frequently supplemented by various non-coding intergenic spacer regions.

In time these regions became so widely used, and indeed so useful, that they represent the animal and plant DNA barcode regions (Hebert *et al.*, 2003, Chase *et al.*, 2007; CBOL Plant Working Group, 2009) – regions of DNA that can be targeted with universal primers for most taxa, amplified and sequenced with ease. These barcodes are then built into large databases enabling amongst others, huge community phylogenetics projects and molecular ecological studies that would otherwise be untenable.

In the field of molecular phylogenetics, researchers moved away from the use of these markers, in part due to features of the markers' evolution and specific ones occurring in their groups of interest (most commonly a lack of adequate variation). The concerted evolution of ITS is still relatively poorly understood, and partial gene conversion can lead to confusing phylogenetic patterns due to the inheritance of different ancestral alleles amongst related taxa of interest. Plastome regions are usually maternally inherited in plants, and at least uniparentally inherited, thereby only revealing aspects of maternal inheritance. The propensity for plants to form polyploids and hybrids meant that investigations of groups containing such species would gain only limited information from these data sources. Introgression of plastids and incongruence between plastid and ITS markers are also issues. As such a move towards low-copy nuclear genes occurred, these being biparentally inherited with the potential to reveal the origins of both parental lineages of hybrids/polyploids, whilst at the same time often (but not always) containing higher amounts of variation compared to plastid markers (Zhang *et al.*, 2012). Complications occur with low-copy nuclear markers, usually due to the difficulty in assigning paralogues or phenomena such as incomplete lineage

sorting (where ancestral polymorphisms were not fixed and this confuses extant patterns of allelic inheritance). The resulting consensus in plant systematics into the 21st century involved a combination of nrITS, plastid and low-copy nuclear genes.

Next-generation sequencing – a plethora of new approaches

With the rise of next-generation sequencing (NGS) or high-throughput sequencing (HTS) techniques, the world of molecular phylogenetics has rapidly begun to change. The first significant change came with the advent of Roche's 454 pyrosequencing, the output of which was several orders of magnitude more data and modest read lengths averaging at best 450 bp. However, the output from 454 sequencing (and the relative cost) ultimately became its downfall, being replaced by Illumina, which is the current world leader in HTS technologies. At the time of writing, Illumina machines can output up to 1 Tb of sequence data (8 billion paired-end reads of 2 x 125 bp) running dual flow cells on a HiSeq with V4 chemistry in 6 days. Whilst this is an enormous leap forward in the amount of sequence data that can be acquired, and a huge reduction in the cost per base, it is still relatively expensive to sequence entire genomes for most species with modest to large genome sizes. It is perhaps *always* unnecessary to sequence entire genomes for systematics applications but at some point in the near future with the advances in sequencing and assembly of genome sequences, this must be an inevitable endpoint.

Notwithstanding the aims and experiments of the broader genomics community, systematists' goals with HTS are somewhat different: the aim is to sequence many taxa whilst retaining as much sequence data per taxon as is currently feasible. This inevitably led to the appropriation of the term 'reduced representation sequencing' to encompass a whole host of strategies to reduce the complexity of the genome (Figure 1.1), and thereby the amount of data

generated for each sample of interest. The most popular approaches are restriction-site associated sequencing, RADseq (Baird *et al.*, 2008; Peterson *et al.*, 2012; Wagner *et al.*, 2013), and target enrichment (baiting) methods (Cronn *et al.*, 2012; Guschanski *et al.*, 2013). The former utilises restriction enzymes (REs) to sequence small regions flanking restriction sites, and by choosing/testing different REs it is possible to change the depth of sequencing per taxon/sample. Target enrichment methods work by first developing bait sequences, usually from other HTS data such as transcriptomes, and choice of the genes/sequences of interest requires much prior thought. Small RNA baits (typically 80-100 nucleotides, tiled at various densities) are then used to essentially fish for DNA of interest, and enrich it using biotin/streptavidin selection and this enriched DNA can then be used for the sequencing run. Both of these methods have shown a great deal of promise in new approaches to molecular systematics, for example RADseq in the Lake Victoria cichlid fish radiation (Wagner *et al.*, 2013) and baits in primates (Guschanski *et al.*, 2013). However, both of these methods require extensive lab procedures and optimisation, unless outsourced to a commercial company, which of course increases the cost; the cost for development of the baits required for target enrichment is also prohibitively expensive for the average systematics or evolution lab.

Genome skimming

The term 'genome skimming' was coined by Straub *et al.,* (2012) as a method by which to 'navigate the tip of the genomic iceberg'. This is, by comparison to other methods, very simple, requiring less lab work and optimisation and in fact no *a priori* knowledge is necessarily needed regarding the organism or indeed its genome size. Genomic DNA is simply extracted and sequenced at low-coverage (perhaps up to 5% of the genome, or 0.05x coverage).

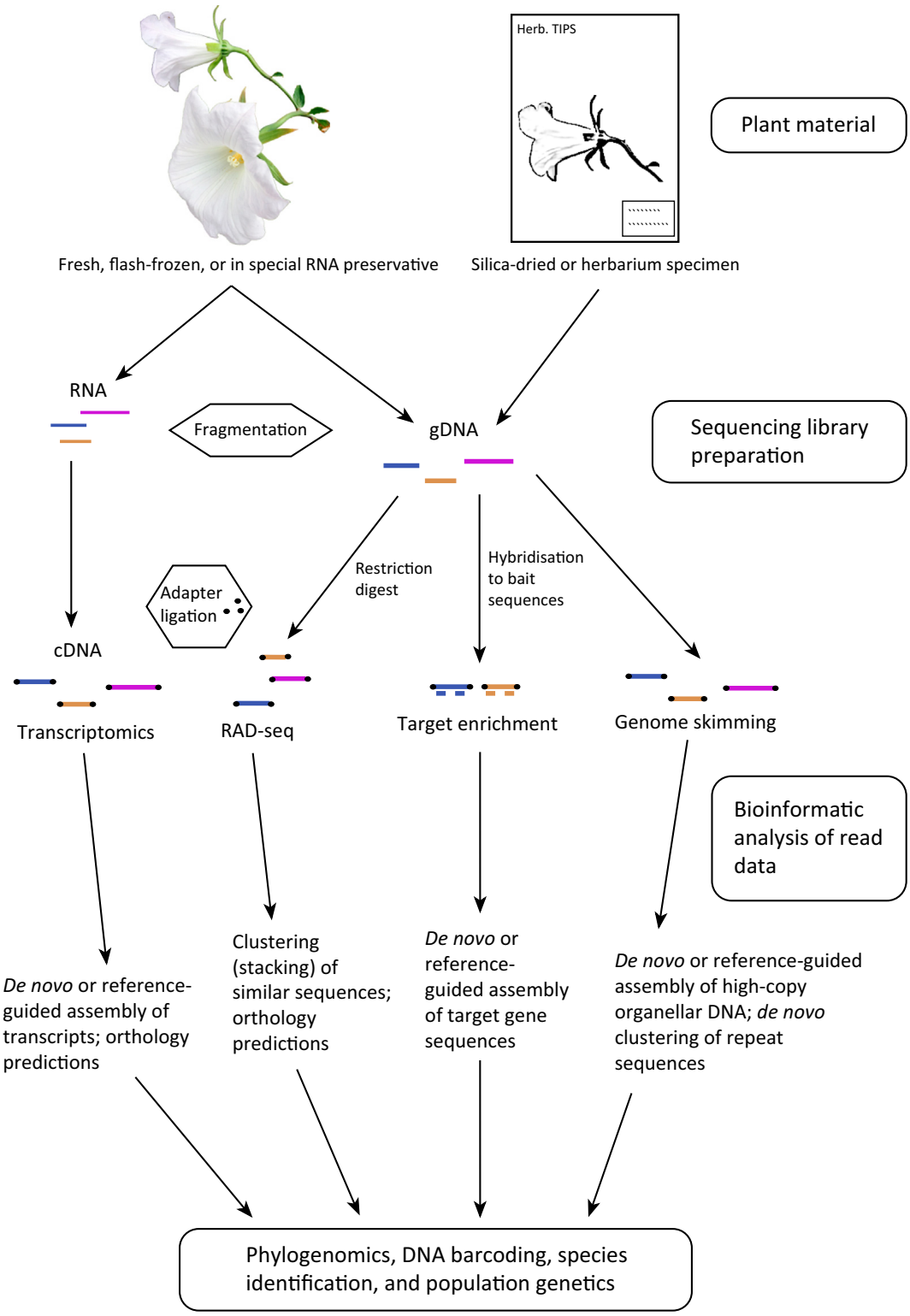


Figure 1.1 A summary of current high-throughput sequencing methods as applied to evolutionary, ecological and systematic studies. Abbreviations: gDNA, genomic DNA; cDNA, complementary DNA. Figure adapted from (Dodsworth, 2015).

Data present in 'genome skims' consists of that which is in high-copy in either the genome or cell, i.e. organellar DNA, high-copy genes, and nuclear genomic repeats (satellites, DNA transposons, retroelements). This method has been widely used to sequence the entire plastome quickly and efficiently, given its usefulness in phylogenetics, at a fraction of the cost (and time) required to do it via conventional Sanger sequencing (e.g. Kane *et al.*, 2012; McPherson *et al.*, 2013; Malé *et al.*, 2014). At the same time as isolating the plastome sequence it is possible to assemble the full rDNA cistron (~7 kb), partial mitogenome sequences, and *de novo* analyse the nuclear repeats in the genome. Full plastome sequences generated by this method have been used to understand the phylogeny of a pantropical tree family (Malé *et al.*, 2014), for phylogeography (McPherson *et al.*, 2013; van der Merwe *et al.*, 2014) and even suggested as a suitable super-barcode to supersede traditional DNA barcodes in plants (Li *et al.*, 2015). The use of whole plastome sequences and the full rDNA cistron represent a natural progression of traditional ITS and plastid markers into the genomic era.

Nuclear genomic repeats in genome skims

An untapped data source in genome skims are the nuclear repeats present in high-copy number in the genomes of most plant species. Repetitive elements consist of tandem repeats (satellites) and transposable elements, of which there are class I (DNA) transposons and class II transposons (LTR and non-LTR retrotransposons). In plants these nuclear repeats can constitute a majority of genomic DNA. The most abundant repeats found in plants are the LTR retroelements of *Ty-1/Copia* and *Ty-3/Gypsy* superfamilies (Hansen and Heslop-Harrison, 2004). Copy number of these repeats is highly variable and as such they contribute a large effect on genome size change. From a systematics standpoint the distribution and occurrence of these repeat types can be phylogenetically informative and provide information about species' evolutionary histories. Traditionally this has been viewed in a cytogenetic

context, via the localisation of repeat types on chromosomes with fluorescence *in-situ* hybridisation (FISH), whereby parsimonious explanations of repeat distributions do reflect the phylogeny of the species (e.g. Lim *et al.*, 2006).

Of course such repetitive elements are present in low-coverage genome skims, along with tandemly duplicated ribosomal DNA. Until recently methods of analysing nuclear repeats were inadequate, with genomics researchers often discarding them in genome assembly projects. A recent pipeline developed to analyse repeats in high-throughput sequencing data has been used successfully to analyse repeats in several groups of plants, including legumes, Solanaceae and Orobanchaceae (Macas *et al.*, 2007; Renny-Byfield *et al.*, 2011; Piednoël *et al.*, 2012). This pipeline utilises graph-based clustering of HTS reads, based on all-to-all BLAST comparisons, and resulting clusters represent *de novo* repeat classes or 'families' (Novak *et al.*, 2010; 2013). In different groups it became apparent that the abundance and distribution of these repeat clusters reflected phylogenetic patterns (Piednoël *et al.*, 2012; Renny-Byfield *et al.*, 2013) and as such represent an underused source of phylogenetic data present in genome skimming datasets.

The genus Nicotiana

Nicotiana L. contains approximately 76 species to date (Knapp *et al.*, 2004), about a third of which are allopolyploids, with several recent homoploid (diploid) hybrids also reported (Clarkson *et al.*, 2010; Kelly *et al.*, 2010; Kelly *et al.*, 2013). Allotetraploids have formed several times in *Nicotiana* over varying timescales, often between the same parental lineages, which sets the genus up as a natural system to unpick the effects of polyploidy over different timeframes (Chase *et al.*, 2003; Clarkson *et al.*, 2004; 2005; Leitch *et al.*, 2008; Table 1.1; and see Chapter 4). This therefore makes *Nicotiana* as excellent genus for studying polyploidy in relation to diversification and adaptation. Polyploids have formed in *Nicotiana*

from as recent as 200,000 years ago (*Nicotiana tabacum* and *N. rustica*) to potentially as old as 10 million years (*Nicotiana* section *Suaveolentes*).

Table 1.1 Comparison of polyploid sections in *Nicotiana*

Section	Age (my)	Paternal parent	Maternal Parent	Taxa
Nicotiana	~<0.2	*N. tomentosiformis*	*N. sylvestris*	1
Rusticae	~<0.2	*N. undulata*	*N. paniculata*	1
Polydicliae	~1	*N. attenuata*	*N. obtusifolia*	2
Repandae	~5	*N. obtusifolia*	*N. sylvestris*	4
Suaveolentes	~10	*N. sylvestris*	hybrid	~26

Nicotiana *section* Suaveolentes

Compared to the rest of *Nicotiana*, section *Suaveolentes* has been the subject of comparatively little study. This is probably due to the remote nature of many of the species in section *Suaveolentes*, most of which are distributed throughout the arid zone of central Australia (Figure 1.2). There have been a smattering of studies focussing on molecular biology, artificial hybridisation and self-incompatibility – and indeed the putative first study of intraspecific variation in plants was conducted in *Nicotiana forsteri* (under the synonym *N. debneyi*) looking at plastid DNA variation (Scowcroft, 1979). Despite this interesting legacy of previous work, the section has since received only a modest amount of effort in terms of molecular phylogenetics, though a doctoral thesis focussing on morphology and species relationships published in 2010 was the most comprehensive piece of work on the section to date (Marks, 2010).

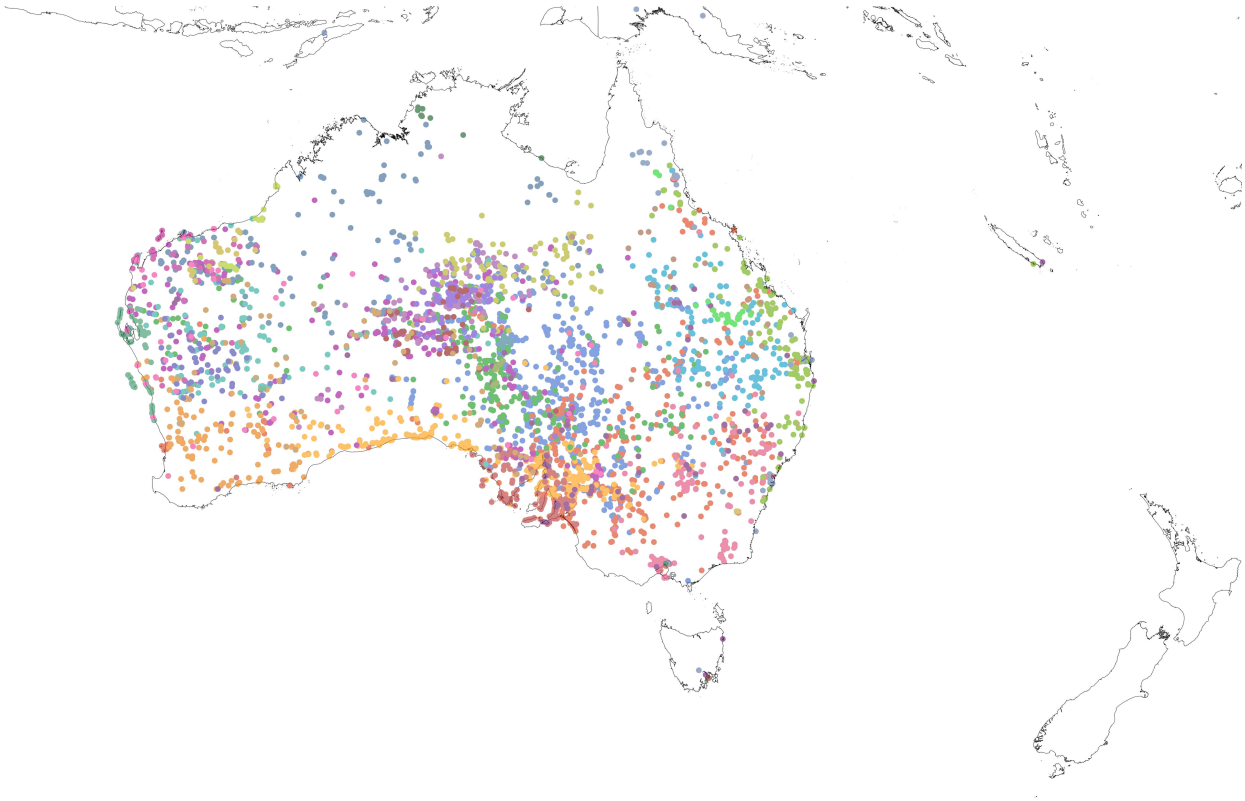


Figure 1.2 Distribution of *Nicotiana* section *Suaveolentes* within Australia. Occurrences are coloured by taxon; all data for identified taxa were downloaded from the Australian Virtual Herbarium.

Section *Suaveolentes* contains approximately 26 taxa, which form a monophyletic group in all previous molecular phylogenetic analyses – based on nrITS, plastid markers, and several low-copy nuclear genes (Chase *et al.*, 2003; Clarkson *et al.*, 2004; 2010; Kelly *et al.*, 2013). The species within section *Suaveolentes* are notoriously homogeneous morphologically, compared with some other clades of *Nicotiana* – and this, in part, may explain why the taxonomy of the group has been considered difficult relative to other sections of *Nicotiana*. However, Marks (2010) following a detailed morphological study of most species in the section, considered the current taxonomy to be stable at the species level and gave detailed descriptions of characters and an updated key to the species of section *Suaveolentes*. For a list of current taxa in the section see Table 1.2 and for examples of morphology see Figure 1.3.

Figure 1.3 Examples of floral and vegetative morphology in *Nicotiana* section *Suaveolentes*. Column A: *Nicotiana simulans* on the Gibber Plains; column B: *Nicotiana truncata* at Fishhole Creek, off the Coober Pedy-Oodnadatta Road.

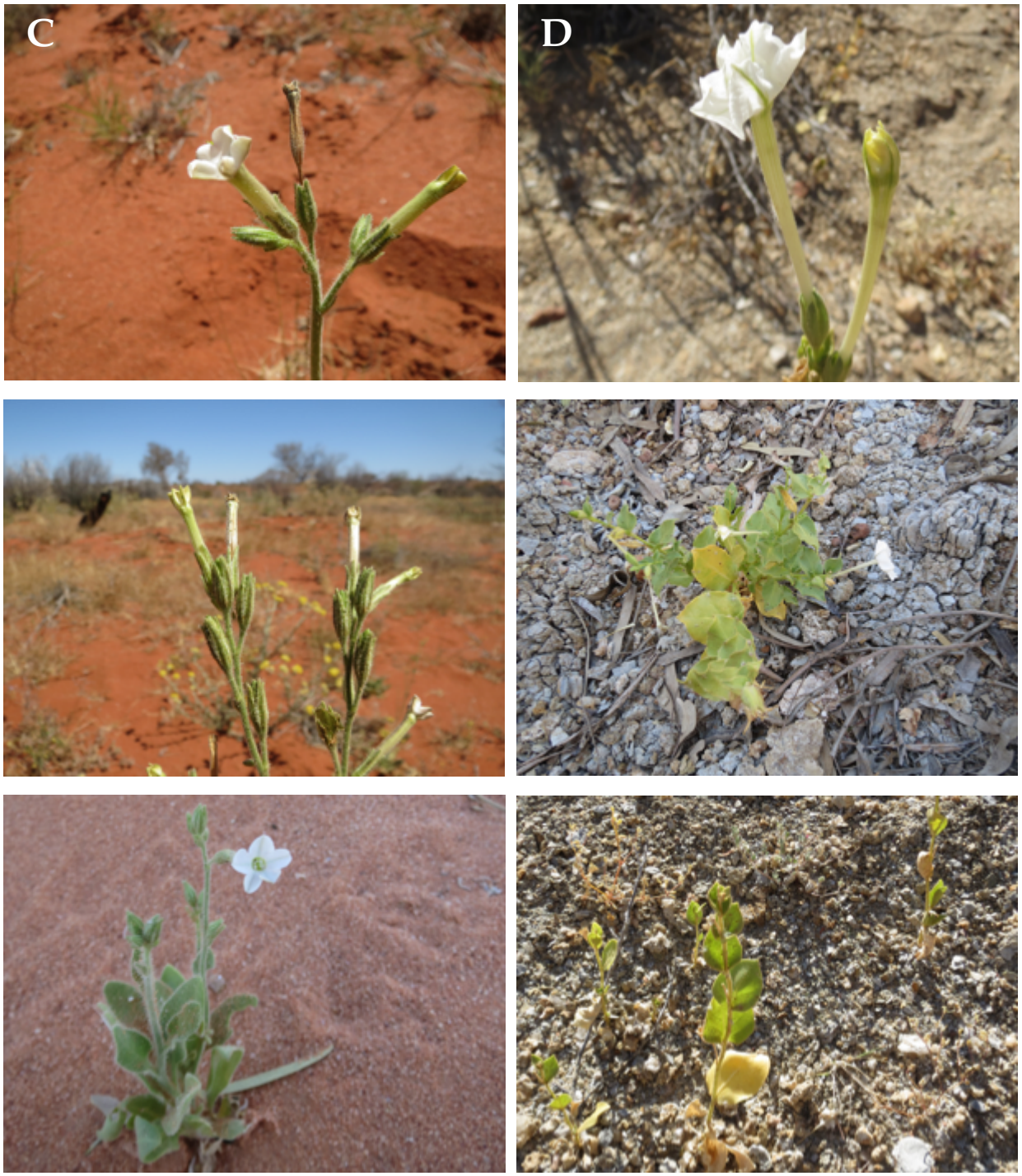


Figure 1.3 *cont.* Column C: *N. velutina* near Oodnadatta; column D: *N. burbidgeae* at Dalhousie Springs, South Australia.

The ancestor of section *Suaveolentes* was an allotetraploid with n = 24, that probably originated within South America ~10 mya (Clarkson *et al.*, 2004; Clarkson, 2007; Leitch *et al.*, 2008; Clarkson *et al.*, 2010; Kelly *et al.*, 2013). This makes section *Suaveolentes* the oldest and most diverse of the polyploid clades within the genus by a long way (for comparison, see Table 1.1). Section *Suaveolentes* has arguably the most complex evolutionary history of *Nicotiana* allopolyploid sections, as the maternal parent of the section is thought to be a homoploid hybrid itself (Kelly *et al.*, 2013). The section also contains a vast range of chromosome numbers from n = 24 down to n = 15 (with n = 14 previously reported for one species, *N. wuttkei*) and only one number never reported in this dysploid series, n = 17 pairs. This heterogeneity in chromosome number post-allopolyploidisation is not found in any other polyploid section, despite in some cases a reasonably large amount of genome size change (e.g. in section *Repandae* with a genome size range of 3.6–5.4 pg despite all species being n = 24).

Nicotiana section *Suaveolentes* represents a recent radiation, and this fact is compounded by previous confusion over its taxonomic circumscription and potentially widespread misidentification of many species in botanic gardens and seedbanks. Up to ~25% of accessions analysed by Marks, (2010) were mislabelled or misidentified. It therefore becomes clear that wild-collected and verified material is essential in order to make inroads into the evolutionary relationships of this group. The intergrading of morphology to some extent, and the phenotypic plasticity of some species (particularly those that are widespread) also add to the difficulty, noted by several authors (Horton, 1981; Marks, 2010). Thus it is also clear that population-level data are needed for as many species as possible, in order to understand the nature of chromosome number change and the phylogenetic relationships of recently diverged lineages within section *Suaveolentes*.

Table 1.2 Currently accepted taxa within *Nicotiana* section *Suaveolentes*, with chromosome number and geographical distribution.

Species	n=	Distribution
N. africana Merxm.	23	Namibia
N. amplexicaulis N.T.Burb.	18	QLD
N. benthamiana Domin	19	WA, NT, QLD
N. burbidgeae Symon	21	SA – Dalhousie Springs
N. cavicola N.T.Burb.	20, 23	WA
N. excelsior (J.M.Black) J.M.Black	19	SA, NT
N. fatuhivensis F.Br.	24/?	Marquesas Islands (South Pacific)
N. forsteri Roem. & Schult.	24	New Caledonia, Lord Howe, NSW, QLD
N. fragrans Hook.	24/?	New Caledonia, Tonga
N. goodspeedii H.Wheeler	16	WA, SA, VIC, NSW
N. gossei Domin	18	NT, SA
N. heterantha Symon & Kenneally	24	WA
N. maritima H.Wheeler	15	SA, VIC
N. megalosiphon Van Heurck & Mull.Arg. subsp. *megalosiphon*	20	QLD, NSW
N. megalosiphon Van Heurck & Mull.Arg. subsp. *sessilifolia* P.Horton	20?	NT, QLD
N. monoschizocarpa (P.Horton) Symon & Lepschi	24	NT
N. occidentalis H.Wheeler subsp. *hesperis* (N.T.Burb.) P.Horton	21	WA
N. occidentalis H.Wheeler subsp. *obliqua* N.T.Burb.	21	WA, NT, SA, NSW, QLD
N. occidentalis H.Wheeler subsp. *occidentalis*	21	WA
N. rosulata (S.Moore) Domin subsp. *ingulba* (J.M.Black) P.Horton	20	NT, WA
N. rosulata (S.Moore) Domin subsp. *rosulata*	20	WA, SA
N. rotundifolia Lindl.	16	WA
N. simulans N.T.Burb.	20	WA, NSW, QLD, NT, SA
N. suaveolens Lehm.	15	NSW, VIC
N. truncata Symon	18	SA
N. umbratica N.T.Burb.	23	WA
N. velutina H.Wheeler	16	NT, SA, QLD, NSW, VIC
N. wuttkei J.R.Clarkson & Symon	16	QLD

Previous molecular phylogenetic studies have included species of section *Suaveolentes*, however they have been limited in their taxonomic sampling (e.g. 13 species in Clarkson *et al.*, 2004; 18 species in Clarkson *et al.*, 2010; 10 species in Kelly *et al.*, 2013). Additionally, aside from low taxon sampling within the section, these studies also analysed a limited amount of variation for these species (Figures 1.4 and 1.5), with phylogenetic analyses typically presenting large polytomies and a lack of a backbone to the tree for section *Suaveolentes*. The phylogenetic hypothesis of Marks (2010) based on a combination of morphology and ITS/glutamine synthase is also largely unresolved (Figure 1.4), but portrays a potential for descending dysploidy as speciation occurred within the section.

Aims and scope of the thesis

High-throughput sequencing and genome skimming have been used for phylogenomics in plants; however, genome skimming has not been tested at shallower phylogenetic depths nor with comprehensive taxon sampling, and most studies have focussed on the use of only a subset of the data present in 'genome skims' – namely organellar DNA. Here I test the usefulness of genome skimming at a low phylogenetic level for phylogenomics of a recent radiation in *Nicotiana*, and develop a means to utilise more of the data – including phylogenetic reconstruction from nuclear genomic repeats, which have been hitherto ignored or discarded by most researchers.

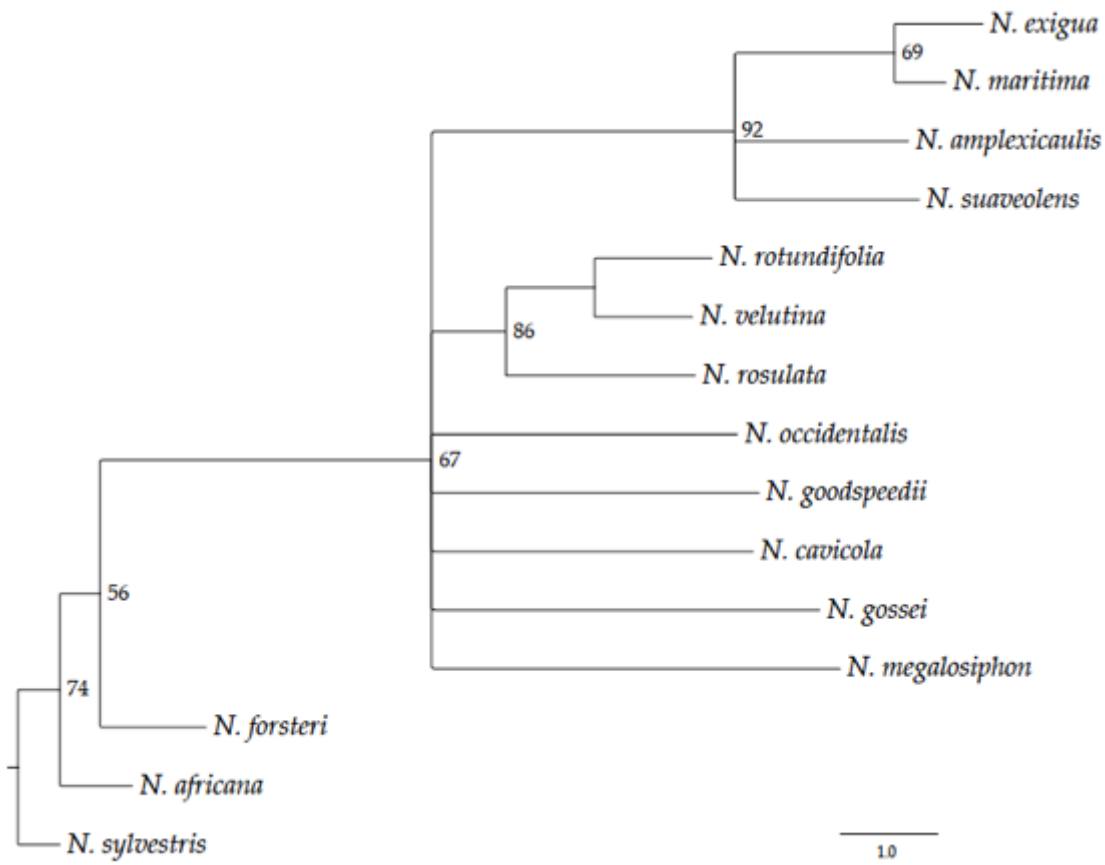


Figure 1.4 Phylogenetic relationships in section *Suaveolentes* based on a re-analysis of the plastid DNA barcode gene, *matK* (sequences taken from Clarkson *et al.*, 2004). Maximum likelihood analysis performed with RAxML under GTR+GAMMA model with 1000 bootstrap replicates, rooted with *N. sylvestris* as the outgroup. Bipartition frequencies are shown on the best scoring ML tree (those >50%).

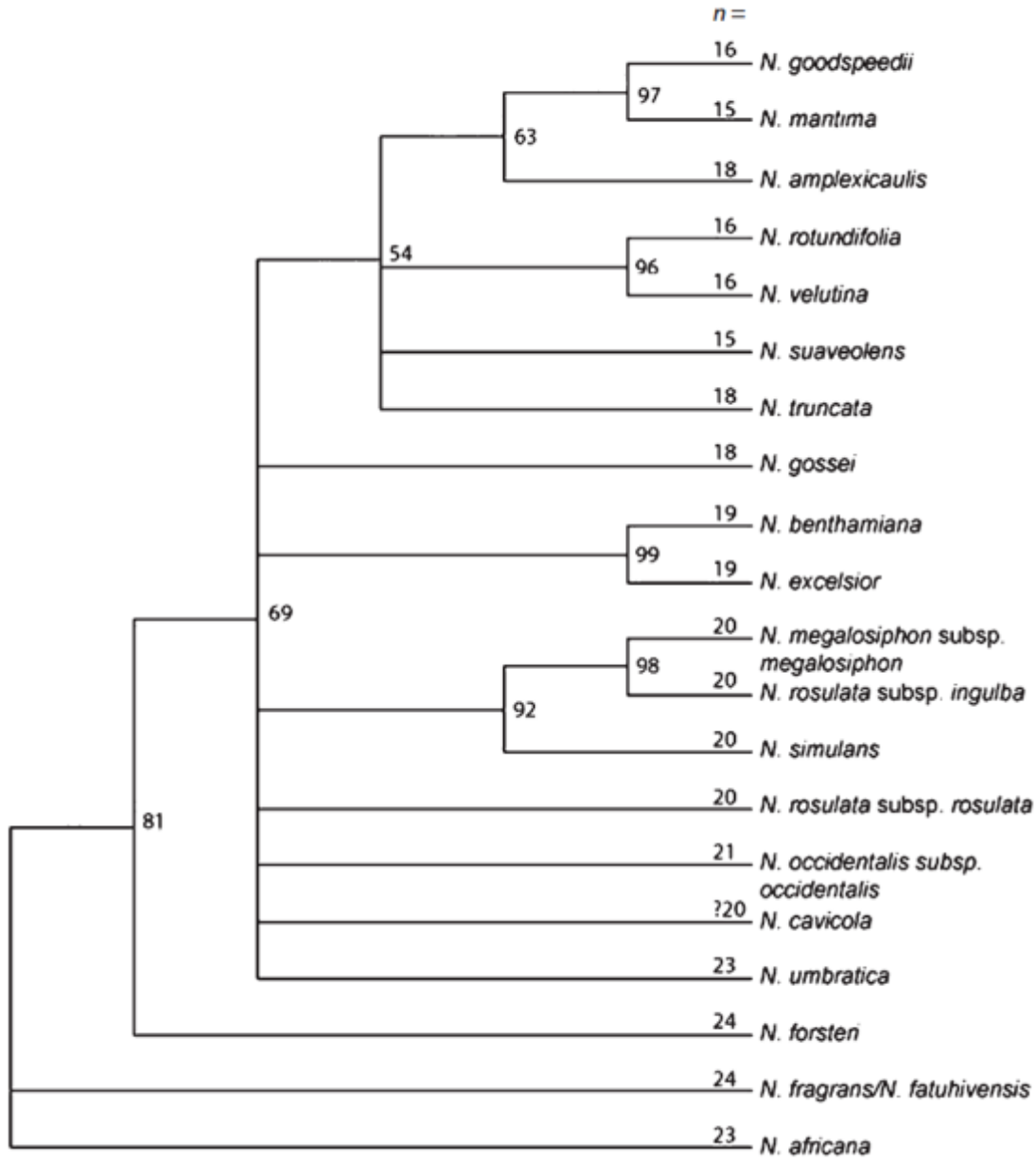


Figure 1.5 Phylogenetic hypothesis for section *Suaveolentes*, adapted from (Marks *et al.*, 2011a). Strict consensus of three equally most parsimonious trees based on a combined dataset of morphology and nuclear-DNA sequences (ITS – Chase *et al.*, 2003; ncpGS long and short copies – Clarkson *et al.*, 2010). Bootstrap values >50% are shown. Chromosome numbers are shown for each taxon at the ends of each terminal branch.

In this thesis I aim to survey the scope of genome skimming in plant phylogenomics, with a focus on utilising nuclear genomic repeats and assembled plastome sequences. This theme runs throughout the thesis. Chapters two and three focus specifically on the development of nuclear repeats for phylogenomics, using graph-based clustering of high-throughput sequencing reads to estimate the abundance of different repeat types, and then using these estimates as continuous characters for phylogenetic inference. The phylogenetic signal in different types of repeats is further explored along with the use of genome skimming data at various phylogenetic levels (chapter 2 focussing more on the intraspecific level).

Chapter four then focuses on *Nicotiana* section *Suaveolentes*: timing of the origin of the section, intrasectional phylogenetic relationships, and the roles of ecological and character evolution in speciation of this group. In this chapter the phylogenetic relationships of section *Suaveolentes* are elucidated with complete taxon sampling and two orders of magnitude more sequence data than previous standard phylogenetic analyses including novel analyses based on repetitive DNA abundances. Additionally, the context of chromosome number change (descending dysploidy) and the nature of genomic and ecological evolution in this recent radiation are investigated using this new phylogenetic framework. The final chapter provides an overview of the results in section *Suaveolentes*, the significance of these for plant and angiosperm evolution in general, and the prospect of genome skimming for plant phylogenomics.

Chapter 2 Genome skimming and nuclear gDNA: Genomic repeat abundances contain phylogenetic signal

Publication information

This chapter is based on the following article, published in Systematic Biology, for which I was the lead author. Co-authors contributed high-throughput sequencing data as follows: Mathieu Piednoël/Susanne Renner (Orobanchaceae); Jiří Macas (Fabaceae); Laura Kelly/Ilia Leitch (*Fritillaria*). All co-authors read, edited and approved the final manuscript.

Dodsworth S, Chase MW, Kelly LJ, Leitch IJ, Macas J, Novák P, Piednoël M, Weiss-Schneeweiss H, Leitch AR. (2015) Genomic Repeat Abundances Contain Phylogenetic Signal. *Systematic Biology*, 64: 112-126.

Summary

A large proportion of genomic information, particularly repetitive elements, is usually ignored when researchers are using next-generation sequencing. Here I demonstrate the usefulness of this repetitive fraction in phylogenetic analyses, utilizing comparative graph-based clustering of next-generation sequence reads, which results in abundance estimates of different classes of genomic repeats. Phylogenetic trees are then inferred based on the genome-wide abundance of different repeat types treated as continuously varying characters; such repeats are scattered across chromosomes and in angiosperms can constitute a majority of nuclear genomic DNA. In six diverse examples, five angiosperms and one insect, this method provides generally well-supported relationships at interspecific and intergeneric levels that agree with results from more standard phylogenetic analyses of commonly used markers. This methodology may prove especially useful in groups where there is little genetic differentiation in standard phylogenetic markers. At the same time as providing data for phylogenetic inference, this method additionally yields a wealth of data for comparative studies of genome evolution.

Introduction

Understanding aspects of comparative evolution, including at its simplest relationships between taxa at varying levels of classification, is being revolutionized by the advent of next-generation sequencing (NGS) technologies. All recent methods in this area are based on multiplexing samples from diverse taxa, thereby maximizing the number of taxa that can be sequenced in one lane or one plate of an NGS run (e.g., Illumina). NGS approaches have enabled a quantum leap in the amount of data available while becoming increasingly cost-effective (Glenn 2011). Approaches include amplicon sequencing (sequencing of specific genes or regions of interest) using barcoded primers (Meyer *et al.*, 2007; Bybee *et al.*, 2011), full mitochondrial and plastid genome sequencing (Timmermans *et al.*, 2010; Barrett *et al.*, 2013; Straub *et al.*, 2012; Kayal *et al.*, 2013), and phylogenomics based on the full complement of protein-coding genes (Zhou *et al.*, 2012; Yoder *et al.*, 2013). Many recent approaches are based on reduced-representation libraries (i.e., reducing genomic complexity/increasing recovery of homologous regions across taxa); in this arena RAD-sequencing based on restriction-site associated DNA fragments scattered across the genome (Rubin *et al.*, 2012; Wagner *et al.*, 2013) and hybridization methods of targeted capture, so-called "pull-down" approaches (Cronn *et al.*, 2012; Carpenter *et al.*, 2013; Guschanski *et al.*, 2013), are two of the most common methodologies.

However, in such phylogenetic/phylogenomics studies, and indeed in broader studies of comparative evolution, the repetitive portion of the genome is often discarded without consideration of any potential use. Repetitive elements in genomes consist of both tandem repeats and interspersed mobile elements (e.g., DNA transposons and retrotransposons). In angiosperms (flowering plants), such repeats are diverse and numerous, contributing up to 70%–80% of nuclear genomic DNA (gDNA), thus making flowering plants an excellent group in

which to study the dynamics of repetitive element evolution (Hansen and Heslop-Harrison 2004; Wicker *et al.,* 2007; Leitch and Leitch 2008; Kelly *et al.,* 2012). Genome sizes vary 2400-fold in angiosperms alone (Pellicer *et al.,* 2010; Kelly and Leitch 2011); aside from cases involving whole genome duplication, much of this variability can be explained by differing amounts of repetitive DNA.

NGS of a small, random sample of the genome (0.5–5% genome proportion [GP], i.e., genome coverage as a percentage) results in data consisting mainly of repetitive sequences; genic regions will not be adequately covered in such a dataset, but repeats present in thousands of copies will be well represented. Previous analyses have shown that low-coverage sequencing of gDNA, followed by graph-based clustering of sequence reads, is sufficient to provide characterization of many hundreds or thousands of well-represented repeats (Macas *et al.*, 2007; Novak *et al.,* 2010; Renny-Byfield *et al.,* 2011); these studies also provide detailed insights into patterns of genome evolution (Renny-Byfield *et al.*, 2011; Leitch and Leitch 2012; Piednoël *et al.,* 2012; Renny-Byfield *et al.,* 2013). Low-coverage gDNA sequencing (i.e., "genome skimming"; Straub *et al.,* 2012) and repeat clustering are now both cost-effective and easy to implement (Novak *et al.,* 2013). The proportion of sequence reads representing a particular repetitive element cluster has also been shown to accurately reflect genomic abundance (Macas *et al.,* 2007; Novak *et al.,* 2010; Renny-Byfield *et al.,* 2011; 2012). Repetitive elements are scattered across the genome and provide much of the characteristic differences between chromosomes and chromosomal subregions, including those in which the majority of genes are embedded (Brookfield 2005). Thus, relative abundance of well-represented repeats is reflective of broad-scale genome composition. Localization of repeats on chromosomes and use of repeats as markers for fluorescence in-situ hybridization (FISH) in some groups has shown that often the most-

parsimonious explanation for these localizations and rearrangements reflect hypotheses of the species tree derived from other data types, usually DNA sequence data (e.g., Lim *et al.*, 2000; 2006). In structure and chromosomal position, repeats in closely related species are nearly identical, whereas more distantly related species diverge in repeat structure and location as genetic similarity decreases.

Here I test the usefulness of a novel phylogenetic methodology based on the abundance of different repetitive elements, estimated through bioinformatic analysis of NGS reads from a small proportion of the genome (Fig. 2.1). Previously similar studies have found that genomic signatures present in the frequency of short sequence repeats can be used to reconstruct phylogenetic relationships, i.e., tetranucleotide frequencies in microbial genomes (Pride *et al.*, 2003) and 2- to 5-nt repeats in birds (Edwards *et al.*, 2002). Here different criteria are used with a clustering method in order to identify homologous repeat classes. This method can essentially be viewed as a hybrid between molecular systematics and morphometric cladistics, as abundances of repetitive DNAs are used as continuously varying characters for phylogenetic inference. I utilize in combination graph-based clustering estimation of repeats (Novak *et al.*, 2010) and the computational methodology of Goloboff *et al.* (2006) in particular, which allows for analysis of continuous characters without assignment (coding) of arbitrarily circumscribed characters, implemented in the software "tree analysis using new technology," TNT (Goloboff *et al.*, 2003a, 2008). Such a combined approach has been utilized successfully with eigenshape-based geometric morphometrics and continuous character phylogenetics in TNT (Smith and Hendricks, 2013). The method is investigated in six diverse groups – five orders of angiosperms and one insect group, with differing genome sizes and amounts of repetitive DNA, and shows a high (but not always identical) level of congruence with previously hypothesized species trees (i.e., current knowledge

from gene trees and morphological circumscriptions) at a variety of taxonomic levels.

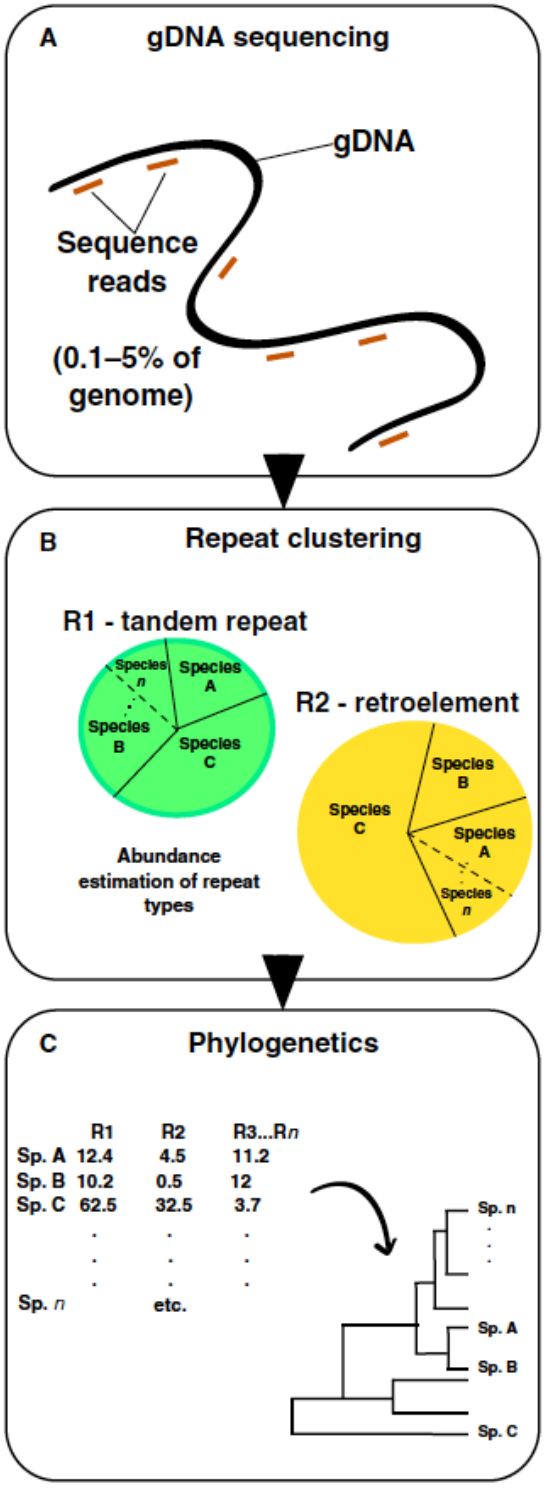


Figure 2.1 Schematic illustrating the workflow for building trees from repetitive DNA abundances. A, low-coverage genomic DNA sequencing using next-generation sequencing methods (NGS; e.g. Illumina). B, clustering of NGS reads using RepeatExplorer pipeline, resulting in abundance estimates of different repeat families. C, phylogenetic analysis in TNT using cluster abundances as continuous phylogenetic characters.

Materials and Methods

Tissue Sources and High-throughput Sequencing of gDNA

Nicotiana—Plant materials (accession numbers), DNA extraction and Illumina sequencing details (including NCBI Short Read Archive [SRA] accession numbers) can be found in Renny-Byfield *et al.* (2012) and Renny-Byfield *et al,* (2013).

Orobanchaceae—Plant materials (including voucher specimen details), DNA extraction and 454 sequencing details (including SRA accession numbers) for this dataset can be found in Piednoël *et al.* (2012). Orobanchaceae is the largest family of parasitic flowering plants. Four genera were included in this dataset, representing a variety of life history strategies: *Lindenbergia*, autotrophic, nonparasitic; *Schwalbea*, parasitic but still photosynthetic; four species of *Orobanche*, nonphotosynthetic, parasitic, including one tetraploid species (*O. gracilis*); and three species of *Phelipanche*, nonphotosynthetic, parasitic.

Fabeae—Seeds of *Vicia tetrasperma* (VIC726), *V. hirsuta* (VIC728), *V. sylvatica* (VIC63), and *V. ervilia* (ERV52) were obtained from the seed bank of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany. Seeds of *Lathyrus sativus* and *L. latifolius* were purchased from Fratelli Ingegnoli S.p.A., Milano, Italy (cat.no. 455) and SEMO Smrzice, Czech Republic (acc.no. 1-0040-68867-01), respectively. *Lathyrus vernus* was collected from a wild population at Vidov, Czech Republic (GPS 48◦55’17.401"N, 14◦29’44.158"E). *Pisum fulvum* (accession ICARDA IG64207) was provided by Petr Smykal, Palacky University, Olomouc, Czech Republic. In all species, genomic DNA was extracted from isolated leaf nuclei (Macas *et al.*, 2007) and sequenced on the Illumina platform (paired-end 100 nt reads) at Elim Biopharmaceuticals, Hayward, USA (P. fulvum) or GATC Biotech, Konstanz, Germany (all other species). Illumina sequencing of *P. sativum* was described in Neumann *et al.* (2012). Voucher

specimens are available for all material sequenced at IPMB, CZ. All read data are available at the SRA with the following accession numbers: *V. hirsuta*—ERR413114; *V. ervilia*—ERR413112; *V. sylvatica*—ERR413113; *V. tetrasperma*—ERR413111; *Lathyrus sativus*—ERR413118 & ERR413119; *L. vernus*—ERR413116 & ERR413117; *L. latifolius*—ERR413120; *Pisum sativum*—ERR063464; *P. fulvum*—ERR413083. Tribe *Fabeae* Rchb. is a group of five genera and ~380 species, containing several important crop species including pea (*Pisum sativum*). In this analysis species from three genera were included, although this includes species proposed to be members of a further two new genera (Schaefer *et al.*, 2012).

Fritillaria—DNA extractions were sourced from the Royal Botanic Gardens, Kew DNA Bank (http://apps.kew.org/dnabank.homepage.html). At the University of Liverpool, 454 sequencing was performed by the Centre for Genomic Research; reads were trimmed to 100 bp prior to clustering, and any reads of <100bp were discarded. All read data are available at the SRA with the following accession numbers: *F. affinis*—ERR571997; *F. alfredae* subsp. *glaucoviridis*—ERR571998; *F. davidii*—ERR571999; *F. imperialis*—ERR572000; *F. koidzumiana*—ERR572001; *F. maximowiczii*—ERR572002; *F. pluriflora*—ERR572003; *F. sewerzowii*—ERR572004; *F. tubiformis*—ERR572005; *Lilium pyrenaicum*—ERR572006. *Fritillaria* is a genus of bulb-bearing petaloid monocots with species possessing some of the largest recorded genome sizes (Ambrozova *et al.*, 2011; Kelly and Leitch, 2011). It comprises approximately 140 species (Rix, 2001) and is closely related to the genus *Lilium*. In this analysis are nine representatives of the genus from each of the two main clades—the North American and the Eurasian clades (Rønsted *et al.*, 2005; Kelly and Leitch 2011).

Drosophila—Illumina reads for the following species were downloaded from the SRA: *Drosophila bipectinata*—SRR345542; *Drosophila suzukii*—SRR1002946;

Drosophila biarmipes—SRR345536; *Drosophila ananassae*—SRR491410; *Drosophila melanogaster*—SRR1005465; *Drosophila sechellia*—SRR869587; *Drosophila simulans*—SRR580369.

Asclepias—Illumina reads for species from the Sonoran Desert clade of *Asclepias* were downloaded from the SRA: *A. macrotis* 149—SRX384308; *A. albicans* x *subulata* 282—SRX384307; *A. cutleri* 382—SRX384306; *A. subulata* 423—SRX384305; *A. macrotis* 150—SRX384304; *A. albicans* 422—SRX384303; *A. subulata* 411—SRX384302; *A. masonii* 154—SRX384301; *A. leptopus* 137—SRX384300; *A. cutleri* 421—SRX384299; *A. coulteri* 45—SRX384298; *A. subaphylla* 272—SRX384297; *A. subaphylla* 271—SRX384296; *A. albicans* 003—SRX384295; *A. syriaca* 4885—SRX040889.

Further details of raw data, quality filtering and resultant read datasets for all examples are provided in Online Appendix 1 (http://dx.doi.org/10.5061/dryad.vn0gc).

Genome Size Estimation

In order to calculate the number of reads for each comparative clustering an accurate genome size should ideally be available for each species. For most datasets genome sizes were available from the Plant DNA C-Values database (http://data.kew.org/cvalues/) or were estimated using flow cytometry. For *Drosophila*, genome sizes were taken from the Animal Genome Size database (http://www.genomesize.com). For *Asclepias* genome sizes were assumed to be equal (423 Mb, as for *A. syriaca*; Bai *et al.*, 2012), as data were unavailable for each species. Without accounting for genome size (e.g., taking the same number of reads), the abundance of each cluster is more likely to reflect genome size rather than the proportion of that repeat in the genome. Genome size ranges (1C) in each dataset are as follows: *Nicotiana* (1.51–5.32 Gb); Orobanchaceae

(0.45–4.32 Gb); *Fabeae* (3.05–9.98 Gb); *Fritillaria* (30.1–75.7 Gb); *Asclepias* (n/a); *Drosophila* (0.16–0.20 Gb).

Clustering of Repetitive DNA

Graph-based clustering of NGS reads was performed as described in Novak *et al.* (2010) using the latest Galaxy based web server implementation of the pipeline, RepeatExplorer (Novak *et al.*, 2013). In brief, all sequence reads (sequence data) are subjected to pair-wise (BLAST) comparison, and similarities are represented by a graph structure in which nodes represent sequence reads and overlapping reads are connected by edges. Edge weights represent the amount of sequence similarity (similarity scores). Clusters of nodes more frequently connected to one another than to outside nodes in the graph represent families of genomic repeats or their parts. Families would be likely to include sequences of the same length (or portions thereof) in which sequence variation is low, 90% similarity over at least 55% of their length.

Combined datasets of reads (sequence data) were compiled as follows: (1) 5% genome proportion (GP) each of four diploid species of *Nicotiana* L. (*N. sylvestris* Spreng., *N. tomentosiformis* Goodsp., *N. attenuata* Torr., *N. obtusifolia* M.Martens & Galeotti); (2) 5% GP each of the four diploid species of *Nicotiana* in (1) and two species of allopolyploid section *Repandae* (*N. repanda* Sims and *N. nudicaulis* S.Watson); (3) 5% GP each of the four diploid species of *Nicotiana* in (1) and two types of *N. tabacum* L. (*N. tabacum* SR1A and *N. tabacum* TR1A synthetic); (4) 2.08% GP each of nine species of Orobanchaceae (*Lindenbergia philippensis* (Cham. and Schltd.) Benth., *Schwalbea americana* L., *Phelipanche ramosa* (L.) Pomel, *Phelipanche purpurea* (Jacq.) Soják, *Phelipanche lavandulacea* Pomel, *Orobanche pancicii* Beck, *Orobanche gracilis* Beck, *Orobanche cumana* Wallr., *Orobanche crenata* Forssk.); (5) 1% GP each of nine species of tribe *Fabeae* (*Vicia sylvatica* L., *Vicia ervilia* Willd., *Vicia hirsuta* (L.) Gray, *Vicia tetrasperma* (L.)

Schreb., *Pisum sativum* L., *Pisum fulvum* Sibth. & Sm., *Lathyrus sativus* L., *Lathyrus vernus* (L.) Bernh., *Lathyrus latifolius* L.); (6) 0.01% GP each of nine species of *Fritillaria* L. and one of *Lilium* L. (*Fritillaria affinis* (Schult. & Schult.f.) Sealy, *F. alfredae* subsp. *glaucoviridis* (Turrill) Rix, *F. davidii* Franch., *F. imperialis* L., *F. koidzumiana* Ohwi, *F. maximowiczii* Freyn, *F. pluriflora* Torr. ex Benth., *F. sewerzowii* Regel, *F. tubiformis* Gren. & Godr., *Lilium pyrenaicum* Gouan); (7) 2% GP each of 15 *Asclepias* L. (*Asclepias syriaca* L. 4885, *A. albicans* S.Watson 003, *A. albicans* 422, *A. coulteri* A. Gray 45, *A. cutleri* Woodson 382, *A. cutleri* 421, *A. leptopus* I. M. Johnst. 137, *A. macrotis* Torr. 149, *A. macrotis* 150, *A. masonii* Woodson 154, *A. subaphylla* Woodson 271, *A. subaphylla* 272, *A. subulata* Decne. 411, *A. subulata* 423, *A. albicans* x *subulata* 282; (8) 5% GP each of 7 *Drosophila* species (*D. ananassae*, *D. bipectinata*, *D. suzukii*, *D. biarmipes*, *D. melanogaster*, *D. sechellia*, *D. simulans*). Different GP values were used across datasets due to genome size differences and the amount of sequencing data available or that could be clustered with the available computing power.

Separate comparative analyses (i.e., simultaneous clustering of reads from all species in the dataset) were run for each dataset on RepeatExplorer (Novak *et al.*, 2013), using default settings (i.e., similarity threshold of 90% over 55% of the read length). Reads were prefixed with codes specific to the taxon in question, enabling comparative analysis of repetitive element abundances in different taxa. Comparative counts of the number of reads in each cluster (which is proportional to their genomic abundance) were used for phylogenetic analyses. Plastid and mitochondrial reads were either filtered out prior to clustering (using BLAST and custom scripts) or were identified after clustering (BLAST to most closely related plastome currently available) and plastid clusters removed prior to phylogenetic inference.

Homology of Repetitive DNA Clusters

The extent to which clusters represent homologous entities is dictated by the similarity parameters specified. The default settings of RepeatExplorer were used, with a threshold similarity of 90% over 55% of the read length to be exceeded in order for a hit to be recorded. Clusters are then produced using a graph-based algorithm and a principle of maximum modularity, which results in clusters where most reads have a high similarity to one another within clusters and a low similarity between clusters (see Novak *et al.,* [2010] for further details on the clustering process). Different repeats will form different clusters in the output of RepeatExplorer and are treated here as separate evolutionary entities (characters). The abundance of a repeat in a species, its genome proportion, depends on repeat copy number and genome size. Tandem repeats have variable monomer sizes up to 180 bp; those with monomer sizes shorter than the read length (typically 100 bp) will form a spherical graph, and those with monomer sizes greater than the read length will form a ring graph structure.

Plant genomes in particular contain a large abundance of LTR retroelements (LTR-REs), which are typically several kb, up to 5kb. These repetitive elements are complex, often dispersed across the genome, and there may be a spectrum of related (or degraded products) of similar LTR retroelements. Based on the RepeatExplorer threshold and graphical algorithm, LTR-REs are often split into different clusters, as parts of these elements are less conserved (e.g., around the LTR) than others (e.g., the protein-coding domains). Sequence divergence within the protein-coding domains is insufficient for phylogenetic analysis, although the number of elements is variable and putatively indicative of evolutionary history. Although LTR-REs may be split over several clusters, they will be split in the same way for every species included in the same clustering run, thereby preserving phylogenetic signal, and each piece of LTR RE would be expected to contain a uniform phylogenetic pattern.

Assembly of High-copy DNA Sequences

High-copy DNA sequences were assembled directly from short read data using the program MIRA (http://www.chevreux.org/projects_mira.html). The general settings used in the manifest file are provided in Online Appendix 1. The following assemblies were performed: (2) *Nicotiana*—large subunit rDNA and whole plastomes were assembled by mapping to *Nicotiana tabacum* sequences as a reference, using raw Illumina reads; (2) *Fritillaria*—whole plastome sequences were assembled directly from plastid 454 reads only (filtered using a custom perl script and BLAST), using the *Lilium longiflorum* plastid genome as a reference; (3) Orobanchaceae—whole plastomes for *O. cumana*, *O. pancicii*, *O. crenata* were assembled from raw 454 reads using the *O. gracilis* plastome as a reference and *P. lavandulacea* was assembled using *P. ramosa* as a reference.

Phylogenetic Analyses

Maximum parsimony analysis

Data matrices consisting of the 1000 most abundant clusters, each representing a repetitive element family, were converted to legal TNT format (modified Hennig86). All abundances were transformed by a constant factor dependent upon the largest cluster abundance in the matrix. Each abundance was divided by this factor (factor = largest abundance/65) in order to make all numbers in the matrices ≤65, the maximal value for continuous character implementation in TNT tree searches (Goloboff *et al.*, 2003a; 2006; 2008). This factorial transformation does not affect the normal distribution of abundance for each cluster and is only necessary for efficient implementation in the TNT program, as described below.

Trees were inferred using maximum parsimony (MP), utilizing the implementation of Farris' algorithm for the down-pass and Goloboff's algorithm for the up pass, as described in Goloboff *et al.* (2006). In such an approach, continuous characters are not arbitrarily recoded but are simply used as additive characters (i.e., count changes can be of noninteger differences). Implicit enumeration (branch and bound) tree searches were used for datasets in this study owing to the small number of taxa in each dataset. Resampling was performed using 100 000 replicates and symmetrical resampling, a modification of the standard bootstrap (Goloboff *et al.,* 2003b). Sequence trees were inferred using the same method for comparison, with gaps coded as missing data. The same phylogenetic reconstruction methodology was employed to enable direct comparison to the repeat trees.

The following datasets were used: (1) full plastomes and 18S-5.8S-26S rDNA for *Nicotiana* diploids and section *Repandae* (assembled); (2) full plastomes for *Fritillaria* (assembled); (3) combined matrix of 17 mitochondrial and nuclear genes for *Drosophila* (28S, *adh*, *amy*, *amr*, *cdc6*, *COI*, *COII*, *ddc*, *esc*, *gpd*, *h2s*, *hb*, ITS, *ND1*, *ND4*, *nup* and *ptc* – see Yang *et al.,* (2012) for GenBank accession numbers); (4) whole plastomes and complete 26S to 18S rDNA cistron for *Asclepias*—alignments taken from Straub *et al.,* (2012); (5) whole plastomes for Orobanchaceae (assembled); and (6) nuclear ITS rDNA and plastid *trnL* from Schaefer *et al.,* (2012) for tribe *Fabeae*.

Maximum likelihood analysis

Maximum likelihood (ML) trees were computed using gene frequency/continuous character implementation in Contml, part of the Phylip package (Felsenstein 1989; 2005). This method assumes that each character evolves independently and only in accordance with random genetic drift, using a Brownian motion model of likelihood. Matrices were transformed such that

cluster abundances represented allele frequencies (0–1) by dividing all clusters by the largest cluster size; this is required for resampling prior to ML tree computation. Resampling from the matrix with replacement (bootstrapping) was first carried out using Seqboot for 1000 replicate datasets. ML analyses were then performed on all 1000 datasets using Contml, and bootstrap percentages mapped onto the strict consensus tree for each dataset computed using Consense.

All trees were viewed in FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and further edited in iDraw (Indeeo, Inc.). All ML trees are shown in Online Appendix 2. Reticulation in the *Nicotiana tabacum* dataset was explored using SplitsTree4 (Huson and Bryant 2006), using 10 000 bootstrap trees from the MP analysis as input for filtered supernetwork analysis (filtering performed at 10% of all input trees). *Nicotiana tabacum* is a relatively recently formed allotetraploid; its two parents have been determined to be *N. tomentosiformis* and *N. sylvestris* (Chase *et al.*, 2003).

Testing Method Performance

To test performance of the method several parameters were analyzed with the smallest clustering dataset (four diploid species of *Nicotiana*) and in TNT as above, but with modifications described below. In each case, the resultant tree was compared with the expected tree topology (Fig. 2.2a) and the symmetric bootstrap percentage recorded.

Reproducibility and relationship with genome proportion

To find the minimal GP necessary to resolve relationships, several sequence datasets were produced at 11 levels of GP from 0.005% to 5.120% (doubling of GP at each step). Three replicate clustering runs were computed for each GP. The mean support and standard error were calculated and used, in addition to tree support and topology, to observe how reproducibility varies with GP.

Relationship between phylogenetic signal and cluster number

To evaluate the number of characters (clusters) sufficient to resolve the tree, trees were built with different numbers of clusters, varying from 5 to 1000 for 25 datasets, and the tree inferred each time, comparing the topology and support percentages.

Variance in phylogenetic signal across the matrix

Variance in phylogenetic signal across the cluster abundance matrix was tested by partitioning it into sets of 150 cluster abundances (this number chosen from the cluster number analysis above). Trees were then inferred from each set, and the resulting resolution and symmetric bootstrap support of the unrooted tree were then estimated. Three GPs were tested (2.00%, 0.32%, and 0.07%) in order to show how different partitions respond at different GPs (chosen to represent difference of three orders of magnitude).

Effect of sampling – range analysis

To evaluate the effect of sampling sequence data on tree building, trees were inferred from clustering of three random samplings of read data. The mean and its standard error were calculated for abundance of each cluster. A phylogenetic analysis based on the range of the mean ±1 standard error of the mean was conducted, as ranges may more accurately reflect the phylogenetic signal in continuous characters (Goloboff *et al.,* 2006), thereby reducing the artefact of two taxa appearing as distinct when they are not. Clusters (repetitive element abundances) that have overlapping normal distributions result in a step count of 0, i.e., no change. Range analysis was tested with a GP of 0.32%.

Phylogenetic informativeness of repeat types

The relative informativeness of different repeat types was analysed by creating subsets of the original matrix based on different repeat annotations. Annotations were assigned to the following categories based on BLAST hits to repbase, custom-protein domain database in the RepeatExplorer pipeline and graph structure: DNA transposon, Ty1/Copia LTR retrotransposon, Ty3/Gypsy LTR retrotransposon, rDNA, satellite, and other (e.g., non-LTR retrotransposon) including unclassified repeats. Matrices were created based on each repeat type, for each example taxon dataset, and trees were inferred as above. The mean bootstrap (as a proxy for tree resolution) was computed for each analysis (Fig. 2.6).

Results

Example 1—Nicotiana (Solanaceae; Solanales)

The unrooted tree presented in Figure 2.2a contains four diploid species of *Nicotiana* and mirrors gene trees based on other nuclear DNA regions (Chase *et al.*, 2003; Clarkson *et al.*, 2004; 2010; Kelly *et al.*, 2013), including rDNA sequences reconstructed from the NGS data (Fig. 2.2). There is a different relationship for these four using the plastome tree constructed from these NGS data (Fig. 2.2a), in line with plastid gene trees published previously (Clarkson *et al.*, 2004).

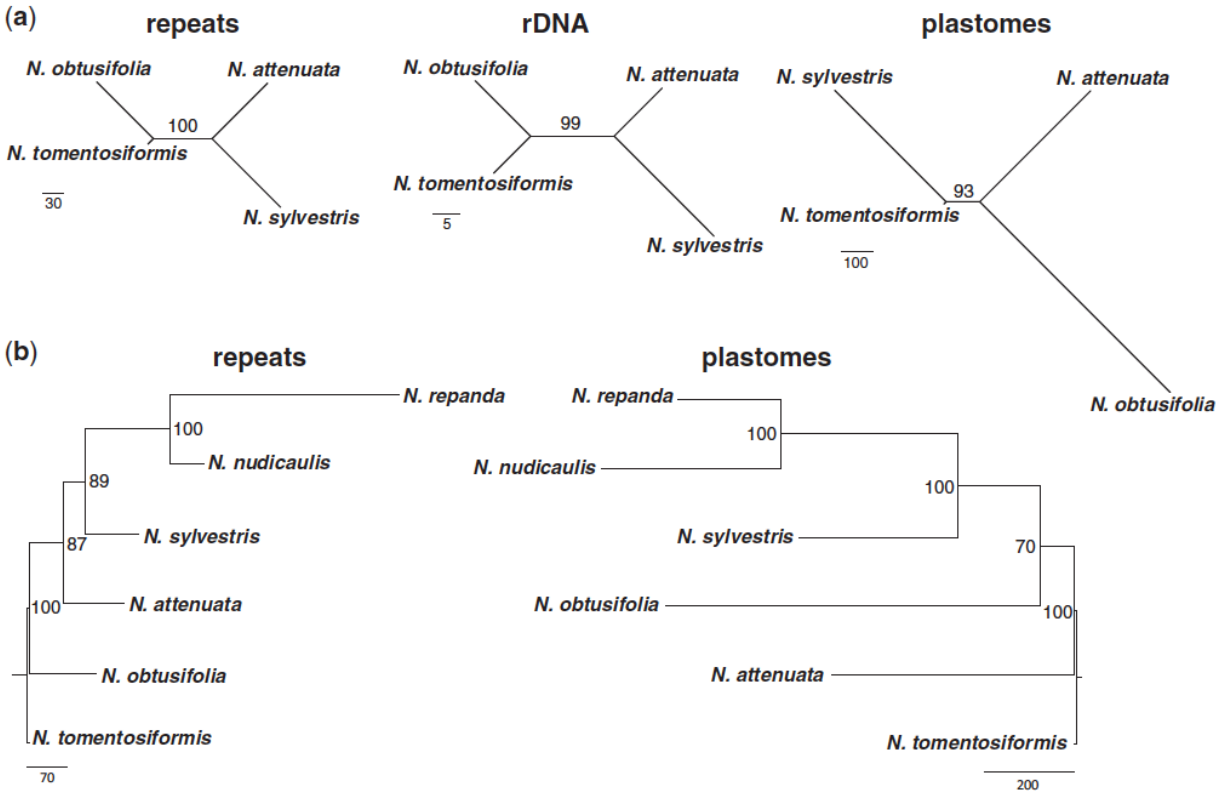


Figure 2.2 Phylogenetic relationships in *Nicotiana* (Solanaceae). a) Unrooted most parsimonious trees for repeats, large rDNA subunit sequences, and plastome sequences for four diploid *Nicotiana* taxa. b) Repeat and plastome trees including diploids from a) and *Nicotiana* section *Repandae* (*N. nudicaulis* and *N. repanda*). Repeat trees are based on 1000 cluster abundances from 5% genome proportion clustering. Maximum parsimony analysis with 10 000 symmetric bootstrap replications and bootstrap percentages plotted onto the single most parsimonious tree in each case. Numbers on nodes represent BPs ≥50; branch lengths are shown from the single MPT and scale bars at the bottom left and right show relative numbers of step changes.

In a further analysis of *Nicotiana*, tree resolution and topology were investigated using data from diploid species and allotetraploid species from two sections of *Nicotiana* (*Nicotiana* sections *Repandae* and *Nicotiana*). *Nicotiana* section *Repandae* is a group of four allopolyploids derived from a single allopolyploid formation event approximately 5 Ma (Clarkson *et al.,* 2005; Parisod *et al.,* 2012; Renny-Byfield *et al.,* 2013). Genomes of sect. *Repandae* have experienced extensive genome turnover subsequent to their formation, and the genomes retain more similarity to the extant relative of their maternal progenitor, *N. sylvestris*, rather than the extant relative of their paternal progenitor, *N. obtusifolia* (Chase *et al.,* 2003; Clarkson *et al.,* 2004; 2005; 2010; Parisod *et al.,* 2012; Kelly *et al.,* 2013; Renny-Byfield *et al.,* 2013). This striking bias is supported here (Fig. 2.2b), where *N. repanda* and *N. nudicaulis* together (100 bp) are strongly supported as sister (89 bp) to *N. sylvestris*. Previous analyses of repetitive DNA and genomic in situ hybridization (GISH) have shown that the genomes of sect. *Repandae* have diverged extensively since their formation, despite being each other's closest relatives, through loss of middle and lower-abundance repetitive elements (Renny-Byfield *et al.,* 2013). This is particularly evident in *N. repanda* (Fig. 2.2b).

To contrast this example, the method performance was investigated with a different allopolyploid section of much more recent origin—*Nicotiana* sect. *Nicotiana*, which contains the familiar allotetraploid *Nicotiana tabacum*, the most common tobacco species in commerce. *Nicotiana tabacum* is estimated to have originated 200 000 years ago or less, and its formation involved entities closely related to extant *N. tomentosiformis* and *N. sylvestris*, its paternal and maternal progenitors, respectively (Chase *et al.,* 2003; Clarkson *et al.,* 2004). GISH is able to distinguish the progenitor genomes (Chase *et al.,* 2003), the T-genome from *N. tomentosiformis* and the S-genome from *N. sylvestris*. Analyses of repetitive DNA show the genome of *N. tabacum* has preferentially lost paternal repeats and is much more similar to *N. sylvestris* (Renny-Byfield *et al.,* 2011; 2012), although its

nrITS and IGS sequences of ribosomal DNA are identical to its paternal progenitor, *N. tomentosiformis* (Chase *et al.,* 2003; Kovarik *et al.,* 2012). In repeat phylogenetic analyses the tree with all four diploid species and *N. tabacum* shows that *N. tabacum* is more closely related to *N. sylvestris* than to *N. tomentosiformis* (Fig. 2.3a), reflecting that abundances of repetitive DNA in *N. tabacum* are in general more similar to those in the maternal parent, *N. sylvestris* (Fig. 2.3a). This result was previously found based on analyses of the GP in different clusters of repetitive DNA, which showed a preferential loss of paternal (i.e., *N. tomentosiformis*) repeats in *N. tabacum* (Renny-Byfield *et al.,* 2011; 2012). Nevertheless, the supernetwork (Fig. 2.3b) illustrates the presence of splits that group *N. tabacum* with *N. tomentosiformis* in addition to placing it with *N. sylvestris*, indicating that some repeats inherited from the paternal progenitor are still present. Analysis of 10 000 bootstrap trees reveals that *N. tabacum* groups with *N. tomentosiformis* in 17% of the trees; in the remaining 83% of trees it is sister to *N. sylvestris*. Additionally, the relatively long branch length separating the *N. tabacum* samples from that of *N. sylvestris* highlights the retention of characters conflicting with its position as sister to *N. sylvestris* (i.e., presence of paternal-type repeats).

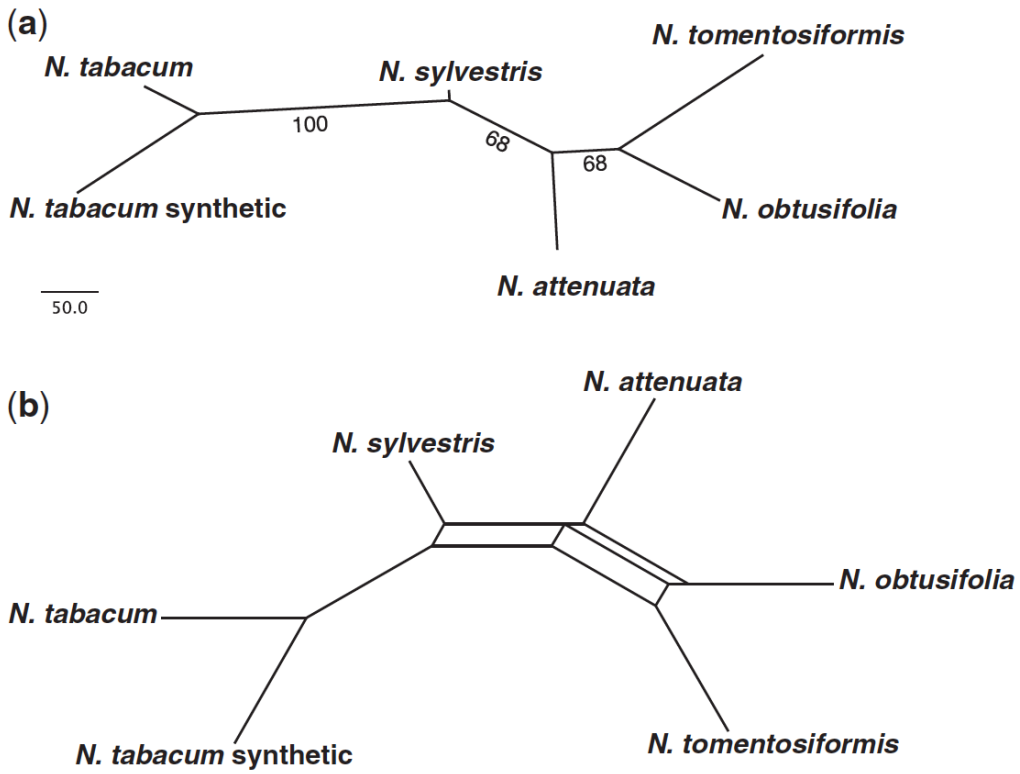


Figure 2.3 Phylogenetic relationships in a young allopolyploid, *Nicotiana* section *Nicotiana* (*N. tabacum*) and related diploid progenitor taxa (Solanaceae). a) Unrooted most parsimonious tree for repeats based on 1000 cluster abundances from 5% genome proportion clustering, maximum parsimony analysis with 10 000 symmetric bootstrap replications and bootstrap percentages plotted onto the single MPT. b) Filtered supernetwork showing relationships present in 10% of the bootstrap trees from a). Numbers on nodes represent BPs≥ 50; branch lengths are shown from the single MPT. The supernetwork is presented in order to present conflicting splits present due to recent reticulation.

Example 2—Fritillaria (Liliaceae; Liliales)

Lilium is the designated outgroup, following Rønsted *et al.* (2005), and the analysis presented here generally places species into their expected clades (Fig. 2.4a). Other than the placement of *F. maximowiczii*, the repeat tree is in agreement with trees based on plastid/plastome data (Fig. 2.4a; Rønsted *et al.*, 2005; Day *et al.*, 2014). The ML tree is partially resolved (Online Appendix 2). Owing to the huge genome sizes in this genus the analysis presented was based on a very low GP of 0.01%, and this may have had an impact on the tree building method—this level of GP should be sufficient, although it is on the cusp of being too low to be representative of repeat diversity and composition (Fig. 2.5a; see discussion below). However, it still reproduces species relationships in a similar manner to previous results, indicating that this number of reads still contains phylogenetic signal despite their low GP.

Example 3—Drosophila (Drosophilidae; Diptera)

Analyses for the diverse fly genus *Drosophila* focused on seven species from the *melanogaster* subgroup. *Drosophila simulans* and *D. sechellia* are strongly supported as sister species, to which *D. melanogaster* is then sister (Fig. 2.4b). *Drosophila suzukii* and *D. biarmipes* form a clade, which is sister to the *D. melanogaster* clade. *Drosophila ananassae* is sister to the rest, with rooting on *D. bipectinata*. These results (Fig. 2.4b) mirror those found in many recent phylogenetic studies based on large amounts of sequence data including mtDNA and nuclear markers (Obbard *et al.*, 2012; Yang *et al.*, 2012; Seetharam and Stuart 2013). The ML analysis mirrors these results with reasonably high levels of support (Online Appendix 2).

Example 4—Asclepias (Apocynaceae; Gentianales)

To test the method on a difficult phylogenetic problem the Sonoran Desert clade (SDC) of *Asclepias* was investigated, and presented alongside the rDNA and

plastome results of Straub *et al.* (2012), with *A. syriaca* as the outgroup. In the repeat tree, *A. macrotis*, *A. coulteri* and *A. leptopus* are supported as separate from the core SDC, which includes *A. albicans*, *A. subulata*, *A. subaphylla*, *A. masonii* and *A. cutleri*; in plastid and nrDNA analyses a strongly supported core SDC excludes *A. cutleri* (Fig. 2.4c). Otherwise the results are different from both rDNA and plastomes, which are in turn different from one another and that from mtDNA (Straub *et al.*, 2012). Note that species are not monophyletic (as with the plastome tree), and a putative homoploid hybrid was included (*A. albicans*, *A. subulata*). Thus this method provides yet another novel hypothesis of relationships between species for this difficult phylogenetic problem, although these relationships are weakly supported. It should be noted however that the results presented may be compromised by a lack of genome size data with which to calibrate the number of input reads. If there are large GS differences between species, then the GP analysed will be significantly different between species, which may influence the topology of the trees produced.

Example 5—Orobanchaceae (Lamiales)

The tree is rooted with *Lindenbergia philippensis* based on the analyses of Park *et al.* (2008) and Piednoël *et al.* (2012). *Orobanche* and *Phelipanche* are each monophyletic (Fig. 2.4d), and *Schwalbea americana* is then sister to them, generally exhibiting similar to greater levels of divergence than analyses based on DNA sequence data (Fig. 2.4d). Such high levels of divergence are only apparent in this dataset. Internal relationships in *Orobanche* and *Phelipanche* are entirely congruent with previous analyses based on nrITS and plastid rps2 sequence data, in all cases with similar or better support in this analysis (Fig. 2.4d; Schneeweiss *et al.*, 2004; Park *et al.*, 2008; Piednoël *et al.*, 2012). The position of *O. gracilis*/*O. cumana* is switched, however, relative to the tree generated from full plastome sequences (Fig. 2.4d). The high support for the position of *O. gracilis* and the fact that there is no evidence of reticulation (result not shown)

are congruent with an autopolyploid origin for this taxon. This could, however, reflect processes of diploidization and genome downsizing in an older allotetraploid, as shown in *Nicotiana* allopolyploids (Example 1); this result could also be affected by low taxon sampling within *Orobanche* (i.e., only one parent present). This might explain the difference between the placement in the repeat analysis and the one based on plastomes.

Example 6—Tribe Fabeae (Fabaceae; Fabales)

Following the tree of Schaefer *et al.* (2012), the tree presented, with section *Ervilia*, now proposed to be genus *Ervilia* (*Vicia hirsuta*, *V. sylvatica*, and *V. ervilia*) as the outgroup, shows a closer relationship between *V. ervilia* and *V. sylvatica*. *Vicia tetrasperma* is the sole representative of section *Ervum* (proposed genus *Ervum*), which is sister to both *Pisum* and *Lathyrus* (Fig. 2.4e). The type section of *Vicia* is not included in this analysis. *Pisum* and *Lathyrus* are sister taxa, and *L. sativus* is sister to *L. vernus* and *L. latifolius*, a grouping that is incongruent with the weakly supported results of the DNA sequence data, in which *L. sativus* and *L. latifolius* are more closely related (Fig. 2.4e; Schaefer *et al.*, 2012).

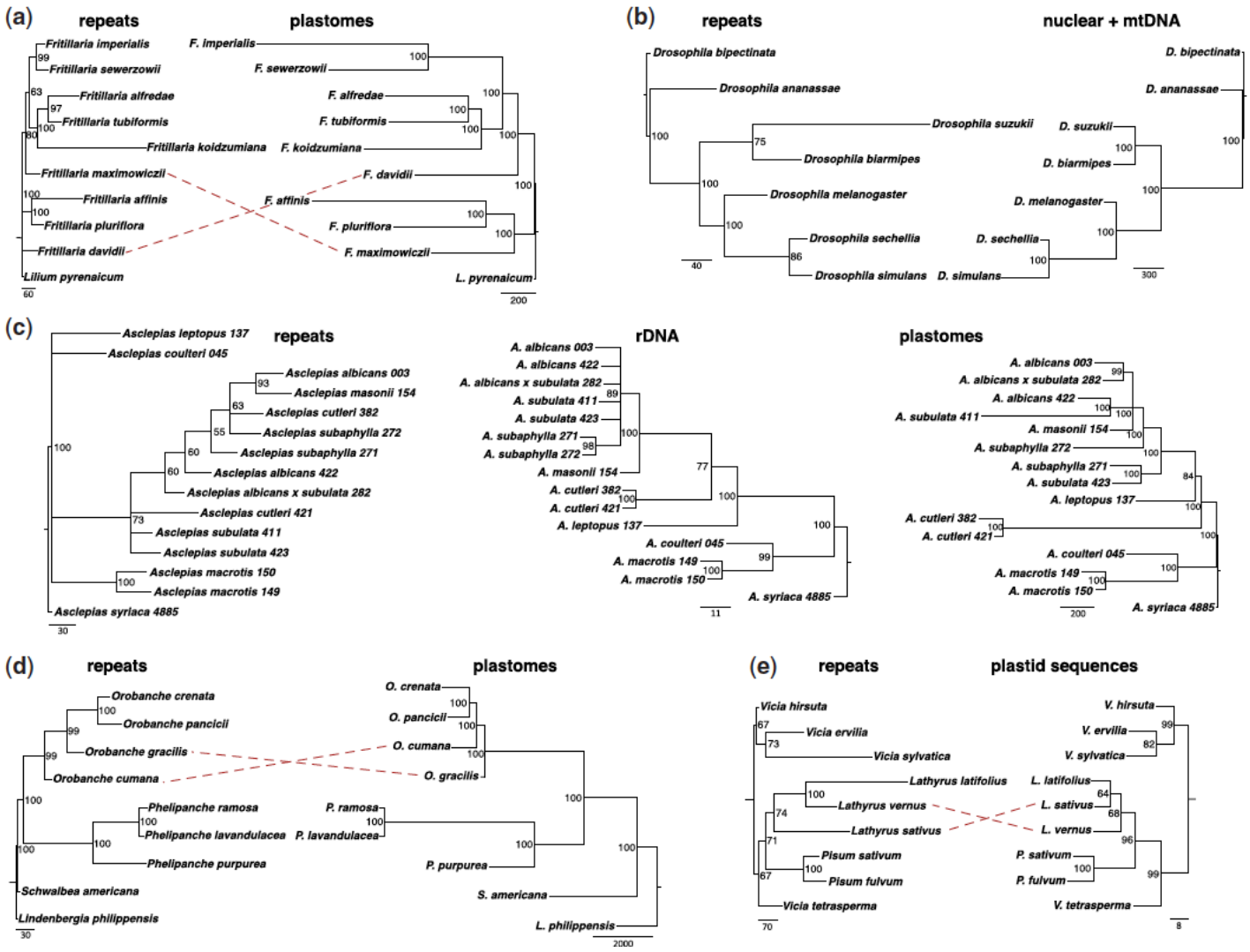


Figure 2.4 Phylogenetic relationships in: a) *Fritillaria* (Liliaceae). Trees for repeats and plastome sequences are shown; repeat tree based on 1000 cluster abundances from 0.01% genome proportion clustering. b) *Drosophila*, the melanogaster species group (Drosophilidae). Trees for repeats and combined matrix of 17 nuclear and mitochondrial genes (see methods for full details); repeat tree based on 1000 cluster abundances from 5% genome proportion clustering. c) The Sonoran Desert clade of *Asclepias* (Apocynaceae). Trees for repeats, 26S to 18S complete rDNA cistron sequences and plastome sequences are shown; repeat tree based on 1000 cluster abundances from 2% genome proportion clustering (assuming the same genome size of 420MBp in each—see methods). d) Orobanchaceae. Repeat tree and plastome tree shown; repeat tree

based on 290 cluster abundances from 2% genome proportion clustering. e) *Fabeae* (Fabaceae). Repeat tree and tree based on combined plastid trnL/nuclear ITS shown; repeat tree based on 1000 cluster abundances from 1% genome proportion clustering. Maximum parsimony analysis with 10 000 symmetric bootstrap replications and bootstrap percentages plotted onto the single most parsimonious tree in each case. Numbers on nodes represent BPs≥50; branch lengths are shown from the single MPT and scale bars at the bottom left and right show relative numbers of changes. Dashed lines show instances of incongruence between repeat trees and DNA sequence trees.

The MP analysis generally has lower support than in other examples, but this broadly mirrors levels of support found in results based on nuclear nrITS and plastid markers (Schaefer *et al.*, 2012). The ML analysis recovers some of the main groupings; however, many nodes collapse, and overall the ML tree is unresolved (Online Appendix 2). An alternative explanation for the lower levels of support found in this example is that perhaps in this case we are looking at a group of more distantly related taxa (comprising several relatively distant related genera), which may be approaching the limits of phylogenetic utility for repetitive DNA. The other examples may well be more closely related, but this is an area in which more investigation is needed to clarify the issues. Compared to the DNA sequence tree, the repeat tree has shorter internal branches and longer external branches, indicating that clustering information may be limited. However deep coalescence and/or extensive reticulation combined with many unsampled taxa may be the underlying problem in this complicated group of legumes. This may explain the poor resolution in Schaefer *et al.* (2012) as well as difficulties in the repeat analyses.

Evaluating Method Performance

Based on resampling the diploid *Nicotiana* data, several aspects of method performance were evaluated. At low GPs, 0.005–0.040%, trees lack resolution, and groupings are often inconsistent with those inferred from sequence data (Fig. 2.5a). Additionally, variance is greater at lower levels of GP, making the method less reliable. Above a GP of 0.1%, clustering and phylogenetic inference appear to be consistent and robust in comparison to trees derived from DNA sequence data. With these GPs, the number of clusters necessary to resolve the tree with high support is approximately 150 (Fig. 2.5b). With lower numbers of clusters (e.g., 5–45), the tree is either topologically inconsistent with trees inferred from sequence data or simply unresolved.

The matrices were then explored using partitions of 150 clusters to test how the phylogenetic signal varied across a large range of cluster abundances, with CL1 being the most abundant (Fig. 2.5c). There is no or little variance in phylogenetic signal across partitions when a suitably high GP has been used (e.g., 2%; Fig. 2.5c). However, when lower GPs are used, the signal degrades rapidly and randomly, particularly with low-abundance elements below cluster number 2000 (Fig. 2.5c). At a GP of 0.07% the trees produced become highly stochastic, either unresolved or showing inconsistent relationships with low support. At a lower GP the signal degrades more quickly as the data essentially become more quickly "coded" into presence/absence characters—the phylogenetic signal in the actual abundance of repetitive elements is entirely lost.

Range analysis showed remarkable similarity between trees produced from three independent samplings of 0.32% (as shown in the GP analysis above). The tree produced from the range of the mean cluster abundance ±1SE of the mean had a topology identical to the three samples as expected, and there are only minor branch length differences between all trees. Thus, there seems to be no significant advantage to undertaking such an analysis, although it may be advisable to do so given the ease of effort and proposed advantage of avoiding spurious groupings where species are not actually statistically different (i.e., they have overlapping normal distributions of cluster abundance).

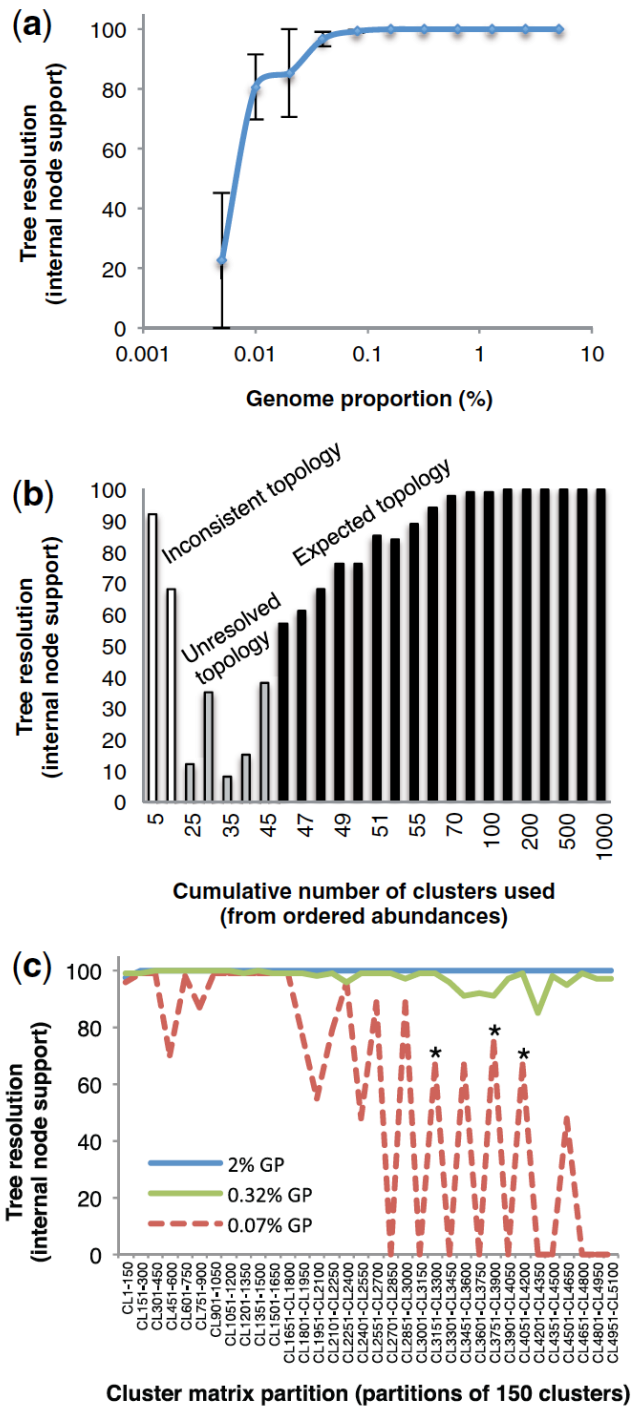


Figure 2.5 Performance measures using the four-taxon diploid *Nicotiana* dataset. a) Analysis of genome proportion (GP%) vs. tree support as the symmetric bootstrap of the unrooted tree. b) Analysis of total number of clusters used vs. tree support as the symmetric bootstrap. c) Partition analysis of 150-cluster segments of the dataset at three levels of GP: 2%, 0.32%, and 0.07%. Asterisks in c) represent trees that contain inconsistent species groupings.

Repeat types differed in their relative informativeness between datasets, but not in a consistent manner. Overall DNA transposons appeared to have less consistent phylogenetic signal than other types of repeats (Fig. 2.6). The relative informativeness of retrotransposons seems to be more taxon-specific, e.g., in some datasets Ty1/Copia were more informative than Ty3/Gypsy, whereas in others the opposite is true. The importance of including unclassified repeats is also highlighted by the informativeness of this category in all datasets examined.

Discussion

Resolving Species Relationships using Repeat Abundances

Using one insect and five angiosperm examples that vary in genome size by ~400-fold (from 1C = 0.2Gb in *Drosophila* to 75.7 Gb in *Fritillaria*), it was shown that using relative abundance of repetitive elements as continuous characters successfully resolves species relationships in a manner similar to that obtained by using DNA sequences from plastid and nuclear ribosomal regions. Using low coverage sequencing of genomic DNA (>0.1%GP) and ≥150 cluster abundances, the repetitive DNA-based phylogeny reconstruction method is consistent in resolving expected relationships (i.e., those produced with other, more standard, methods, and data). The method can be seen as an additional source of phylogenetic information from the repetitive, noncoding portion of the genome, which will be a useful comparison to results based on DNA sequences. Gene trees represent the ancestry of particular sequences, and because allele histories may have differences from the species trees plus have coalescent times that differ from each other and species divergence times (Pillon *et al.*, 2013) they often present conflicting topologies (Nichols, 2001). Thus, there has been a recent focus on multiple sequence datasets, which aim to give a genome-wide assessment of species divergence.

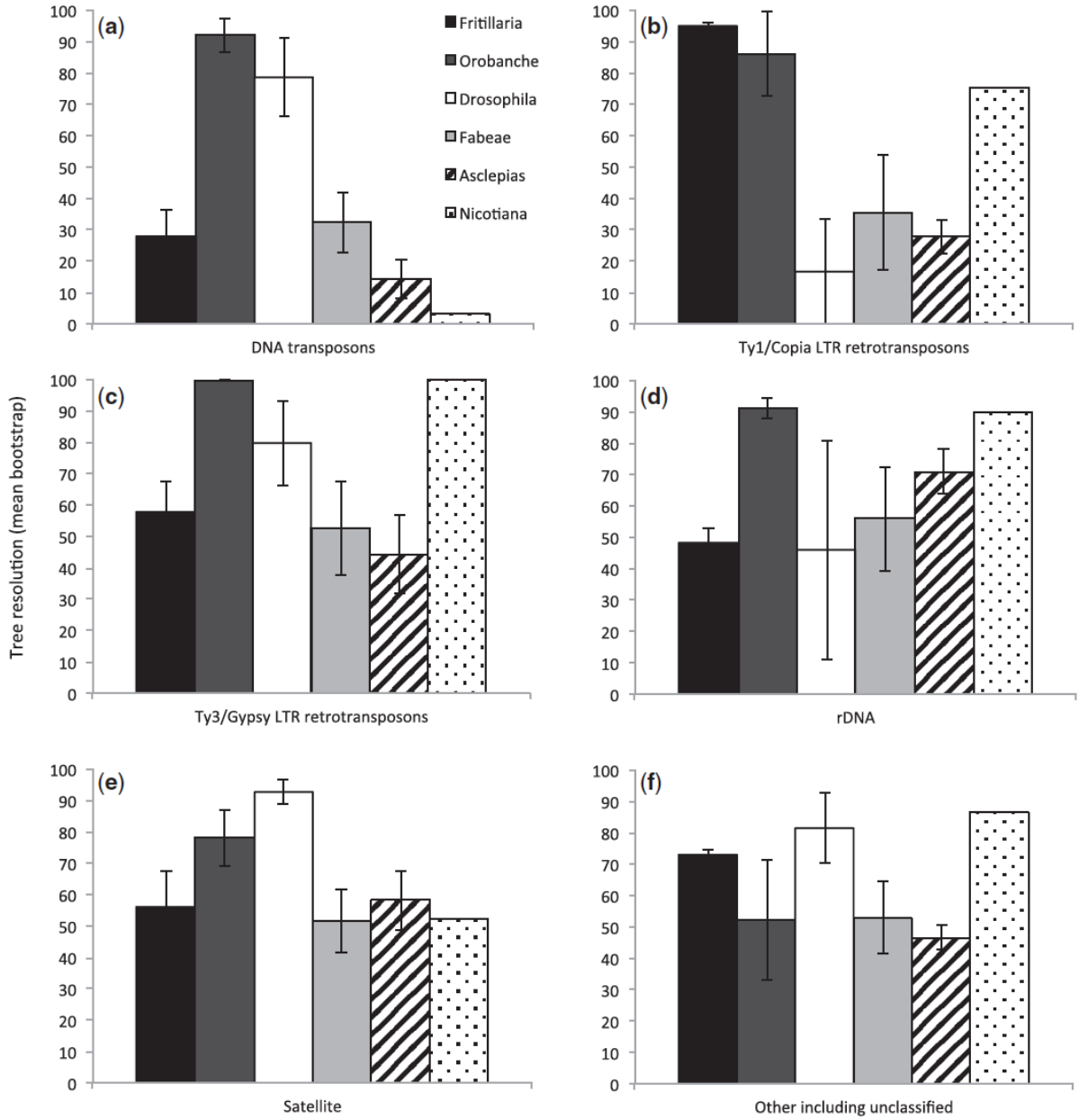


Figure 2.6 Impact of repeat type on tree resolution and method performance. Informativeness of each repeat type was estimated by creating subsets of the original matrices based on repeat annotation; in each case the mean bootstrap was calculated for each repeat type and each taxon dataset, error bars represent the standard error. a) DNA transposons. b) Ty1/Copia LTR retrotransposons. c) Ty3/Gypsy LTR retrotransposons. d) rDNA. e) Satellites. f) Other repeats including unclassified repeats and non-LTR retrotransposons.

Here, the repeat method provides data widespread across the genome, and each repeat abundance (cluster) is a marker; a matrix of such abundances likely represents many independently evolving characters. Furthermore at an appropriate level of GP there is little variance in phylogenetic signal across the dataset, showing remarkably consistency, although this requires further testing with additional datasets that include a larger number of taxa.

ML results on the whole agree with maximum parsimony results, although for many groups the tree is only partially resolved, e.g., *Fritillaria* and tribe *Fabeae* (Online Appendix 2). ML does not perform as well as MP here, which may be due to the large state space of the character coding might provide insufficient information to accurately inform site likelihoods. Further likelihood and Bayesian models for inferring trees from repeat abundances are being developed, but these approaches are limited by our understanding of repeat evolution, which is in its infancy. It is possible that horizontal transfer occurs for several repeats, but the overall impact of this on the results is believed to be low, due to the large proportion of the genome covered by these analyses. Additional work is needed to further model the evolution of repeat populations, but these results and others (e.g., Jurka *et al.*, 2011; 2012) provide evidence that repeats evolve primarily in accordance with random genetic drift; they therefore contain useful phylogenetic signal.

For reliable estimates of species relationships, it is recommended to use >0.1%GP for clustering and subsequent phylogenetic analysis based on at least 150 cluster abundances. If a lower GP is used it is suggested to use 1000 cluster abundances in the analysis for reliable detection of the phylogenetic signal present. With larger (and more repetitive genomes) it is possible to resolve relationships with lower levels of GP as read depth will likely still be sufficient at GP<0.1%, as observed here for *Fritillaria*.

Why use Repeat Abundances for Phylogenetics?

There is an extensive literature on the evolution of repeats, see e.g., recent reviews by Kelly *et al.* (2012), Leitch and Leitch (2012), Kejnovsky *et al.* (2009), and Fedoroff (2012). All genomes contain tandem repeats and transposable elements, predominantly retrotransposons and derivative repeats—in plants these are typically long-terminal repeat (LTR) Ty1/Copia and Ty3/Gypsy elements (Hansen and Heslop-Harrison 2004). Copy number of these elements is highly variable and can change rapidly, contributing a large effect on genome size and architecture (Kelly and Leitch 2011). Genome content of these repeats, as a whole is the result of mechanisms of repeat expansion (e.g., retrotransposition, repeat recombination) and contraction (e.g., recombination-based deletion). In this study it is shown that the variable abundance of different repetitive elements contains phylogenetic signal, i.e., one reflecting the evolutionary history of these species, which would be expected, given the premise that repeats are an inherent structural feature of the genome and in fact underlie much of the evolution of large, complex eukaryotic genomes (Fedoroff 2012).

Previously, it has been shown that the frequency of short sequence repeats provides genomic signatures that can be used to reconstruct phylogenetic relationships (e.g., Edwards *et al.,* 2002; Pride *et al.,* 2003). This approach builds on this insight to show that a genomic signature containing phylogenetic information extends to many larger repeat classes (Fig. 2.6), based on sequence similarity and graphical clustering. Potentially therefore it is likely to be robust to particular features of individual genomes being analysed, which may or may not be rich in certain categories of repeat. Researchers using genome skimming approaches to assemble high-copy DNA features (plastomes, rDNA cistron) already have these repeat data available (e.g., Bock *et al.,* 2014), and it provides

another data source of noncoding nuclear DNA from which to infer phylogenetic hypotheses.

This method is genome-wide, and there is no need to attempt to distinguish paralogues, as essentially each cluster represents a homologous family of repeats (as determined by their highly conserved sequence, which is how they are clustered in the first place), the genomic abundance of which is used as phylogenetic data. In contrast, using low-copy nuclear markers requires paralogues to be distinguished accurately from one another. Low-copy nuclear genes are becoming increasingly popular (Zhang *et al*., 2012) due to often-higher variability when compared with plastid or mitochondrial markers (but not always, Turner *et al.,* 2013), particularly for recent radiations and population studies. High variability can be seen as a consequence of having long coalescent times (compared to plastid or mitochondrial markers), but this in fact confounds their use, as when two or more young taxa share ancient alleles only due to the absence of fixation in ancestral populations rather than unique descent (Pillon *et al.*, 2013). This means that many such markers are needed to provide evidence for which ones are providing spurious or conflicting results. The repeat method, however, shows particular promise for these cases, as repeats evolve rapidly, and there is neither the added complication of comparing markers with different coalescent times nor effect of longer coalescent times and incomplete lineage sorting.

This method proved useful for inferring parents of some allopolyploid taxa in which extensive diploidization has occurred (i.e., replacement of repeats typical of one parent with those of the other) and repeat abundance is much more similar to one of the parents (here observed in groups of *Nicotiana* allopolyploids). In other groups, in which homoploid and polyploid hybridization occurs, further conflict will occur in construction of strictly

bifurcating trees, but this can be analysed more thoroughly using networks or pruning of putative parental or hybrid taxa. Using each cluster abundance (a single character) to make a separate estimate and then building a network based on these input trees, it may be possible to extract in the repeat data further evidence of reticulation.

Nonetheless, in these analyses repeat abundances have proven useful for determining one of the putative parents of allopolyploid taxa (Renny-Byfield *et al.*, 2013), but perhaps with some refinement in methods it could be expected to demonstrate clear evidence for both parents, especially in recently synthesized hybrids (as in the *N. tabacum* analyses when investigating the individual trees from the bootstrap replicates). Additionally, this method has not yet been tested in a species or species complex of wide geographical range, which would be a useful further test for how this method performs at the intraspecific level. If homogenization of repeats occurs as a result of gene flow, which holds back formation of sequence variants, the prediction is that divergence occurs quickly once gene flow ceases, as might occur for isolated populations of a widespread species.

There is an additional caveat that should be mentioned; in order for the method to work reliably it is best that genome size is first estimated (via flow cytometry). If genome size is not used to standardize repeat abundances, then the resultant trees may reflect genome size differences more than shared evolutionary history, as repeat estimates reflect sampling bias rather than true abundance. It should be noted that the *Asclepias* example did not include genome size standardization as these data were not available. Additionally, lack of genome size information was simulated in the *Fritillaria* example, using the same number of reads (40 000) for each sample, despite >2-fold genome size differences (30–75 Gb), but results from this analysis closely mirror those

presented in the results section, again with high levels of support (Online Appendix 3). Thus when genome sizes are unknown, the method still has the potential to produce a reliable phylogenetic result. It should therefore be possible to include data from various genomic sources, including available genome sequence data, although this may vary relative to sequencing effort/coverage used in the clustering.

Further Applications and Development of this Method

The proposed method may also prove useful for sequencing DNA from herbarium specimens, where genomic DNA is often degraded to a greater extent (Särkiinen *et al.*, 2012). Highly degraded DNA will be expected to contain intact copies of high-copy regions (i.e., particularly shorter repeats) even when other regions are largely eliminated. Herbarium specimens provide an invaluable source of plant material, often collected from remote regions and for taxa that may have since become rare or even extinct. Utilizing this resource will be a continued focus of research in plant systematics, and bridging the gap from Sanger sequencing of short markers to NGS of genome-wide markers is one current focus of research. Here, one possible solution to this problem is provided, as short NGS reads of low-coverage gDNA give us an invaluable insight into repetitive DNA proportions, a method that is advantageous because repeats are present in high-copy number and distributed across the genome. Unlocking and mining data from museum collections will understandably be a future focus of systematic studies (e.g., Guschanski *et al.*, 2013).

The clustering pipeline utilized here has been shown to be effective in characterizing repetitive elements across various groups of eukaryotes including, for instance, bats (Pagan *et al.*, 2012). Phylogenetic trees based on repeat abundances estimated with RepeatExplorer could be produced in other groups of animals, including mammals, and fungi. This makes the methodology

particularly useful for those carrying out genome evolutionary studies in various types of organisms.

Concluding Remarks

In conclusion, this methodology is quick and easy to implement from the initial stage of DNA extraction and Illumina sequencing through to clustering of reads and tree building, utilizing the RepeatExplorer pipeline and TNT program (both freely available). As the cost of NGS continues to decrease in coming years (currently the major cost is in the library preparation), the overall cost of using this method will also decrease.

This method has proven successful in resolving species relationships as previously hypothesized by analyses of DNA sequence data (plastid and nuclear trees) and morphological circumscription in five diverse groups of angiosperms and one insect. There were only a few instances where the results are incongruent with trees derived from DNA sequences, for which evidence about cause should be sought; there are good reasons why in some cases plastid DNA and rDNA could be misleading, so these discrepancies should not be ignored or be thought of as a fault of using repeats as phylogenetic characters. This method does provide an important extension of molecular systematics methods and should be useful for comparative phylogenomics. At the same time as providing data for robust phylogenetic reconstruction in diploid species, this method provides abundant information for understanding genome evolution in the context of repetitive DNA. Indeed, this has already been done for a number of the datasets/partial datasets used in this study (e.g., Piednoël *et al.*, 2012, 2013; Renny-Byfield *et al.*, 2013).

Chapter 3 Using genomic repeats for phylogenomics: A case study at the intraspecific level

Publication information

This chapter is based on the following article, published in the Biological Journal of the Linnean Society, for which I was the lead and corresponding author. All co-authors read, edited and approved the final manuscript.

Dodsworth S, Chase MW, Särkinen T, Knapp S, Leitch AR. (2015) Using genomic repeats for phylogenomics: A case study in wild tomatoes (*Solanum* section *Lycopersicon*; Solanaceae). *Biological Journal of the Linnean Society*, DOI: 10.1111/bij.12612

Summary

High-throughput sequencing data have transformed molecular phylogenetics and a plethora of phylogenomic approaches are now readily available. Shotgun sequencing at low genome coverage is a common approach for isolating high-copy DNA, such as the plastid or mitochondrial genomes, and ribosomal DNA. These sequence data, however, are also rich in repetitive elements that are often discarded. Such data include a variety of repeats present throughout the nuclear genome in high copy number. It has recently been shown that the abundance of repetitive elements has phylogenetic signal and can be used as a continuous character to infer tree topologies. In the present study, repetitive DNA data in tomatoes (*Solanum* section *Lycopersicon*) is evaluated in order to explore how they perform at the inter- and intraspecific levels, utilizing the available data from the 100 Tomato Genome Sequencing Consortium. The results add to previous examples from angiosperms where genomic repeats have been used to resolve phylogenetic relationships at varying taxonomic levels. Future prospects now include the use of genomic repeats for population-level analyses and phylogeography, as well as potentially for DNA barcoding.

Introduction

One of the simplest approaches to using high-throughput sequencing for phylogenetics is to randomly sequence a small proportion of total genomic DNA. The sequences of reads present in these datasets are biased towards sequences with the greatest numbers of copies in the genome (Straub *et al.*, 2012); this includes not only high-copy organellar DNA, such as the plastid and mitochondrial genomes, but also ribosomal DNA and the many kinds of repeats, particularly retrotransposon sequences (Dodsworth *et al.*, 2015a). Molecular systematics relies on the alignment of homologous DNA sequences, whether coding or noncoding, and subsequent phylogenetic trees are inferred based on patterns of differences in these alignments. Repetitive elements are not suitable for such analyses in exactly the same way. For example, although retrotransposons have homologous protein domains involved in element mobility, the sequence divergence of these domains between taxa is not sufficient to resolve phylogenetic relationships. What does vary, and in many cases drastically, is the abundance of particular retrotransposons and other repeat types. This abundance of homologous repeats can then be used as a quantitative character for phylogenetic reconstruction.

Recent tools have been developed that allow us to analyse, quickly and efficiently, the repetitive portion of the genome from low-coverage genome sequencing data, and then to use these data for phylogenetic inference (Novak, *et al.*, 2010; Novak *et al.*, 2013; Dodsworth *et al.*, 2015a). This methodology has been shown to be effective for inferring phylogenetic relationships in well-studied groups of angiosperms in several different families (Apocynaceae, Fabaceae, Liliaceae, Orobanchaceae, and Solanaceae). Typically, the method does not work well above the level of genus because there are often no repeats in common (and therefore no shared characters on which to infer phylogenetic relationships). Understanding how repetitive elements could be used in

phylogeographical and population genetic studies, as well as in resolving difficult phylogenetic problems at the species-level, is now a focus for future research.

In the present study, the usefulness and power of nuclear repeat regions is tested at inter- and intraspecific levels. This is performed using wild and cultivated tomato species, including multiple cultivars as a case study to explore intraspecific variation in genomic repeats and the subsequent performance of these datasets in phylogenetic inference. The wild tomatoes present an excellent case study as a result of the availability of genomic and genetic data, and extensive previous analyses of phylogenetic relationships using plastid markers, low-copy nuclear markers, nuclear ribosomal internal transcribed spacers, and amplified fragment length polymorphisms (Peralta *et al.*, 2008; Grandillo *et al.*, 2011). Four informal groups are recognized within the section: (1) 'Lycopersicon group' with *Solanum lycopersicum*, *Solanum cheesmaniae*, *Solanum galapagense*, and *Solanum pimpinellifolium* (the 'red/orange fruit' clade); (2) 'Arcanum group' with *Solanum arcanum*, *Solanum chmielewskii*, and *Solanum neorickii* (the 'green fruit' clade); (3) 'Eriopersicon group' with *Solanum huaylasense*, *Solanum chilense*, *Solanum corneliomulleri*, *Solanum peruvianum*, and *Solanum habrochaites*; and (4) 'Neolycopersicon group' containing only *Solanum pennellii*, which was considered to be sister to the rest of the section by (Peralta *et al.*, 2008) based on its lack of the sterile anther appendage that occurs as a morphological synapomorphy in *S. habrochaites* and the rest of the core tomatoes.

More recent studies using conserved orthologous sequence markers (COSII; Rodriguez *et al.*, 2009) and genome-wide single nucleotide polymorphisms (SNPs) (Aflitos *et al.*, 2014; Lin *et al.*, 2014) have largely supported previous hypotheses with respect to major clades within the tomatoes, although

individual species relationships are less clear cut for some taxa. The extent to which multiple evolutionary histories can be recovered through the analysis of different genomic fractions has been explored in tomatoes, and concordance analysis revealed significant discordance possibly as a result of biological processes such as hybridization or incomplete lineage sorting (Rodriguez *et al.*, 2009). There are no reported polyploids in this clade, and there is not much variation in genome size, although macro- and microgenome rearrangements are reported (Tang *et al.*, 2008; Szinay *et al.*, 2010, 2012; Verlaan *et al.*, 2011).

Materials and Methods

Taxa sampled

Sampled material from 20 accessions included all currently recognized species of the core tomato clade (section *Lycopersicon*) and *Solanum tuberosum* L. (potato) as the outgroup. Representatives of *Solanum* sect. *Lycopersicon* (Table 1) (Peralta, Knapp & Spooner, 2005; Peralta *et al.*, 2008) included in the analyses were: *S. lycopersicum* L., *S. arcanum* Peralta, *S. corneliomulleri* J.F. Macbr., *S. cheesmaniae* (L.Riley) Fosberg, *S. chilense* (Dunal) Reiche, *S. chmielewskii* (C.M.Rick, Kesicki, Fobes & M.Holle) D.M.Spooner, G.J.Anderson & R.K.Jansen, *S. galapagense* S.C.Darwin & Peralta, *S. habrochaites* S.Knapp & D.M.Spooner, *S. huaylasense* Peralta, *S. neorickii* D.M.Spooner, G.J.Anderson & R.K.Jansen, *S. pennellii* Correll, *S. peruvianum* L., and *S. pimpinellifolium* L. Seven accessions representing different cultivars of *S. lycopersicum* were also included.

High-throughput sequence data acquisition

Illumina sequence data from the 100 Tomato Genome Sequencing Consortium (Aflitos *et al.*, 2014) were downloaded from the NCBI Short Read Archive (SRA), with the accession numbers: ERR418040 – *S. lycopersicum* 'Alisa Craig' LA2838A; ERR418039 – *S. lycopersicum* 'Moneymaker' LA2706; ERR418048 – *S. lycopersicum* 'Sonata' LYC1969; ERR418055 – *S. lycopersicum* 'Large Pink' EA01049;

ERR418056 – *S. lycopersicum* LYC3153; ERR418058 – *S. lycopersicum* PI129097; ERR418078 – *S. lycopersicum* LYC2962; ERR418093 – *S. arcanum* LA2172; ERR418061 – *S. corneliomulleri* LA0118; ERR418087 – *S. cheesmaniae* LA0483; ERR418098 – *S. chilense* CGN15530; ERR418085 – *S. chmielewskii* LA2663; ERR418121 – *S. galapagense* LA1044; ERR410244 – *S. habrochaites* LYC4; ERR418096 – *S. huaylasense* LA1365; ERR418091 – *S. neorickii* LA0735; ERR410253 – *S. pennellii* LA716; ERR418084 – *S. peruvianum* LA1278; and ERR418082 – *S. pimpinellifolium* LA1584. 454 sequence data for the outgroup *Solanum tuberosum* (ERR023045) were also downloaded from the SRA because appropriate Illumina data were unavailable. There are different sequencing biases based on 454 or Illumina technologies (and library preparation protocols) and, ideally, they should not be mixed; however, the outgroup has been clearly defined based on extensive literature and therefore any difference in this one taxon should not have any impact upon the ingroup taxa results.

Dataset preparation and subsampling of read data

SRA files were unpacked into FASTQ using the FASTQ-DUMP executable from the SRA Toolkit. FASTQ files were then filtered with a minimum quality of 10 and converted to FASTA files. For the 454 data, reads were trimmed to 100 bp and filtered. All samples were assumed to have a genome size of approximately 1 Gb based on data available on the Plant C-values Database that shows little variation in genome size between species within section *Lycopersicon* (831–1198 Mbp) (Table 1) (http://data.kew.org/cvalues). Each accession was then sampled for 0.2% of the genome by randomly subsampling each Illumina/454 dataset. This resulted in 20 000 reads of 100 bp per sample from all *Solanum* accessions. The reads in each sample were labelled with a unique nine-character prefix, making a total combined dataset of 400 000 reads. In addition, a further dataset was compiled to test the above assumption that genome size is comparable between species of section *Lycopersicon*. The 20 taxa were randomly

shuffled and half were down-sampled to 14 000 reads, representing 0.7 of the original sample. This proportion was chosen because it reflects the genome size variation currently found within the section (approximately 831/1198).

Table 3.1 Taxa sampled including accession details, short read archive accession number for genomic data, and genome size (http://data.kew.org/cvalues)

Species	Accession	Cultivar	Short Read Archive accession number	Genome size (1C – Mbp)
Solanum arcanum	LA2172	NA	ERR418093	1125*
Solanum cheesmaniae	LA0483	NA	ERR418087	905
Solanum chilense	CGN15530	NA	ERR418098	1125*
Solanum chmielewskii	LA2663	NA	ERR418085	NA
Solanum corneliomulleri	LA0118	NA	ERR418061	NA
Solanum galapagense	LA1044	NA	ERR418121	905–1002*
Solanum habrochaites	LYC4	NA	ERR410244	905
Solanum huaylasense	LA1365	NA	ERR418096	1125*
Solanum lycopersicum	LYC2962	NA	ERR418078	1002
Solanum lycopersicum	PI129097	NA	ERR418058	–
Solanum lycopersicum	LYC3153	NA	ERR418056	–
Solanum lycopersicum	LYC1969	Sonata	ERR418048	–
Solanum lycopersicum	LA2706	Moneymaker	ERR418039	–
Solanum lycopersicum	LA2838A	Alisa Craig	ERR418040	–
Solanum lycopersicum	EA01049	Large Pink	ERR418055	–
Solanum neorickii	LA0735	NA	ERR418091	NA
Solanum pennellii	LA716	NA	ERR410253	1198
Solanum peruvianum	LA1278	NA	ERR418084	1125
S. pimpinellifolium	LA1584	NA	ERR418082	831
S. tuberosum	DH Kuba 48/6	NA	ERR023045	856

*Values assumed based on previous intraspecific status within other taxa. NA, not available.

Clustering analysis using RepeatExplorer (RE)

Clustering of Illumina/454 reads was performed using the RE pipeline, implemented in a Galaxy server environment (http://www.repeatexplorer.org) as described in Dodsworth *et al*. (2015a). RE clustering was used to identify genomic repeat clusters within each dataset, with default settings (minimum overlap = 55 and cluster size threshold = 0.01%). Briefly, using a BLAST threshold of 90% similarity over 55% of the read length, RE identifies similarities between all sequence reads and then identifies clusters based on a principle of maximum modularity. To identify and discard any potential plastid repeat clusters, the *S. lycopersicum* plastid genome (HG975525.1) was used as a custom repeat database. Plastid repeats are not considered informative in a phylogenetic context because their high abundance is likely linked to the dynamics of photosynthesis in different tissue types and species, and therefore is not indicative of evolutionary history. Hence, plastid regions need to be identified prior to using genomic repeat data in phylogenetic analyses. In this study, none of the clusters were identified as belonging to the plastid genome and hence no regions were removed. Finally, RE was used to identify the 1000 most abundant repeats for phylogenetic analyses, as measured by read numbers per cluster.

Phylogenetic analysis using cluster abundances

The top 1000 most abundant clusters were used to create a matrix for phylogenetic inference. Cluster abundances were used as input characters. To make the cluster abundance values smaller based on requirements of input data for TNT, all abundances were divided by a factor of 18.5 (= largest cluster abundance/65) so that all data would fall within the range 0–65 (as required by the TNT software). Tree topologies were inferred using maximum parsimony as implemented in the TNT software with continuous character states enabled (Goloboff *et al*., 2006; Goloboff, Farris & Nixon, 2008) following settings in

Dodsworth *et al*. (2015a). Continuous characters are not recoded in any way and are used as 'normal' additive characters, except that count changes can now be non-integer differences (i.e. numerical). Tree searches were performed using implicit enumeration (branch- and bound) with 10 000 symmetric bootstrap (BS) replicates. To explore reticulation in the dataset, a network approach was employed. SPLITSTREE4 (Huson and Bryant, 2006) was used to create a filtered supernetwork from the 10 000 bootstrap trees from the maximum parsimony analysis, with filtering set at 10% of all input trees (i.e. 1000 trees).

Results

Phylogenetic relationships in Solanum sect. Lycopersicon

The single most parsimonious tree from the analysis of genomic repeats recovers *S. habrochaites* and *S. pennellii* as the first branching taxa within section Lycopersicon (Fig. 3.1). The 'Eriopersicon group' (sensu Peralta *et al*., 2008; *S. corneliomulleri*, *S. peruvianum*, *S. huaylasense*, and *S. chilense*) is recovered with high branch support (100 BS). *Solanum neorickii* ('Arcanum group') is recovered as sister to all remaining species (99 BS) (Fig. 3.1). Members of the 'Arcanum group' (*S. chmielewskii* and *S. arcanum*) are found to be nested within the clade consisting of all the members of the 'Lycopersicon group' (Fig. 3.1). The results do not recover a red–orange fruited clade but do find three of the species bearing red or orange coloured fruits in a strongly supported clade (*S. lycopersicum* LA2838A, *S. pimpinellifolium* and *S. galapagense*; 86 BS) within a polytomy including all other red- and orange-fruited accessions.

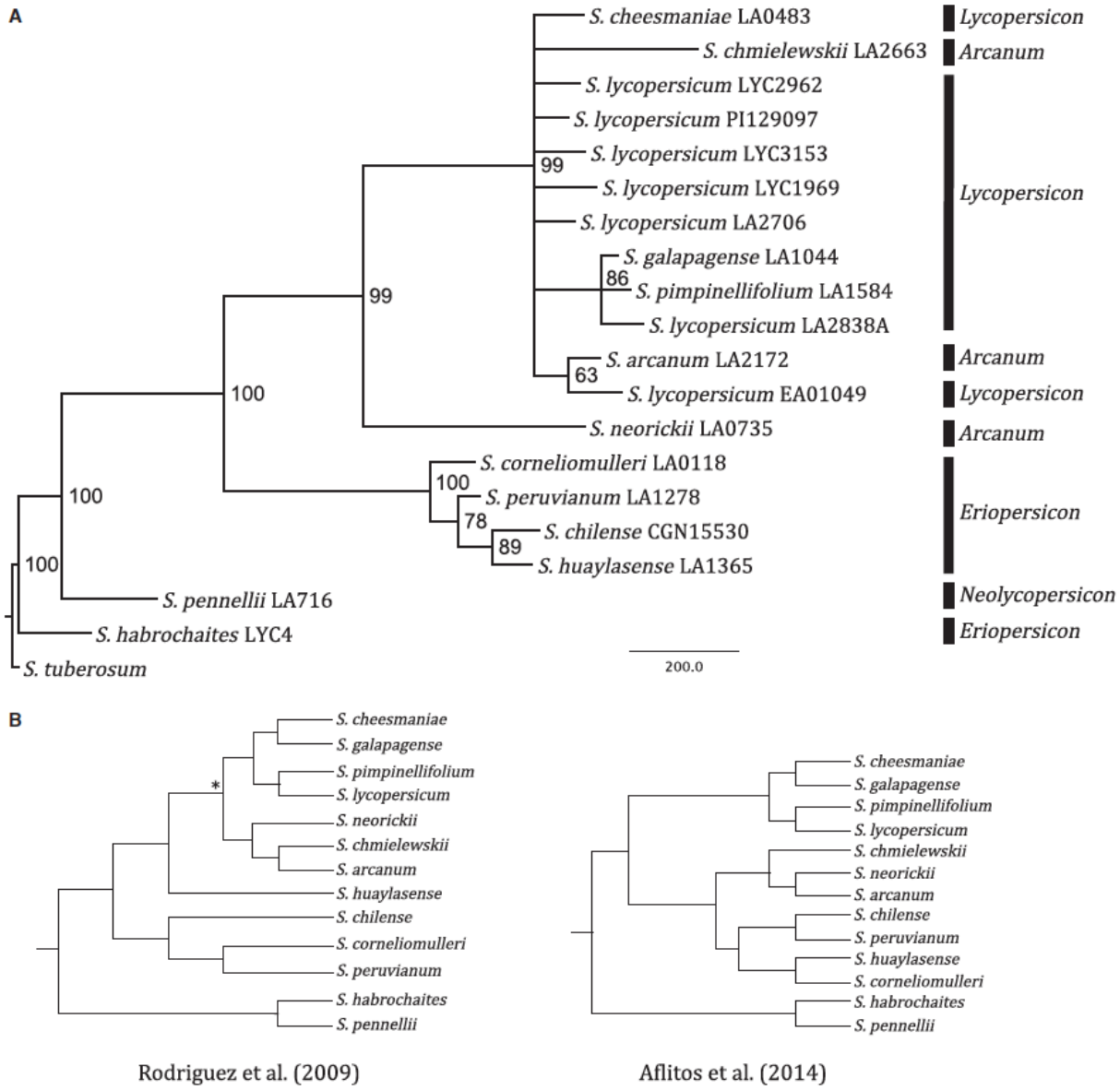


Figure 3.1 Phylogenetic relationships in *Solanum* section *Lycopersicon*. A, the single most parsimonious tree topology is shown based on abundance values of the 1000 most abundant repeats identified in Illumina/454 next-generation sequencing runs. A total of 0.2% of the genome for each accession was used. Bootstrap values are shown for each node (10 000 symmetric bootstrap replicates). Branch lengths are proportional to numerical step changes in repeat abundances (scale bar). Accession numbers are given for each sample. Current taxonomic grouping is indicated according to informal groups *sensu* Peralta *et al*. (2008). B, summarized phylogenetic hypotheses from Rodriguez *et al*. (2009) and Aflitos *et al*. (2014); low support is indicated by asterisks.

The additional analysis testing the effect of genome size variation on tree inference produced the same overall phylogenetic results consistent with those based on equal sampling of 20 000 reads (data not shown), although there are some differences in the large clade containing all *S. lycopersicum* accessions. However, this clade is still largely unresolved. Network analyses show evidence of reticulation in this clade, as indicated by the presence of splits present in the filtered supernetwork (Fig. 3.2).

Each accession had a unique combination of repeat percentages, as reflected in the difference in terminal branch lengths. Some accessions also had unique repeat types not found in any other accession (Fig. 3.3); the largest numbers of unique repeats were found in *S. habrochaites* and *S. pennellii*, with 239 and 301 clusters, respectively, out of the 1000 most abundant clusters. One accession of cultivated *Solanum lycopersicum* (EA01049) had one unique repeat type (Fig. 3.3).

Discussion

Relationships in Solanum section Lycopersicon

The taxonomy and estimates of phylogenetic relationships within the core tomato clade (*Solanum* section *Lycopersicon s.s.*) have begun to be stabilized in recent years (Peralta *et al.*, 2005; 2008) using a variety of different markers from both the plastid and nuclear genomes. Rodriguez *et al.* (2009) used a suite of COSII nuclear markers to identify five strongly supported clades within the broader tomato group (incl. sections *Juglandifolia* and *Lycopersicoides*). Their results supported monophyly of section *Lycopersicon* as treated in the present study, and did not resolve either the position of their strongly supported *S. arcanum+S. chmielewskii+S. neorickii* or the relationships of these species with each other. The data presented here show a similar lack of resolution regarding

these three taxa and, additionally, place them within a large polytomy including all the red- and orange-fruited taxa.

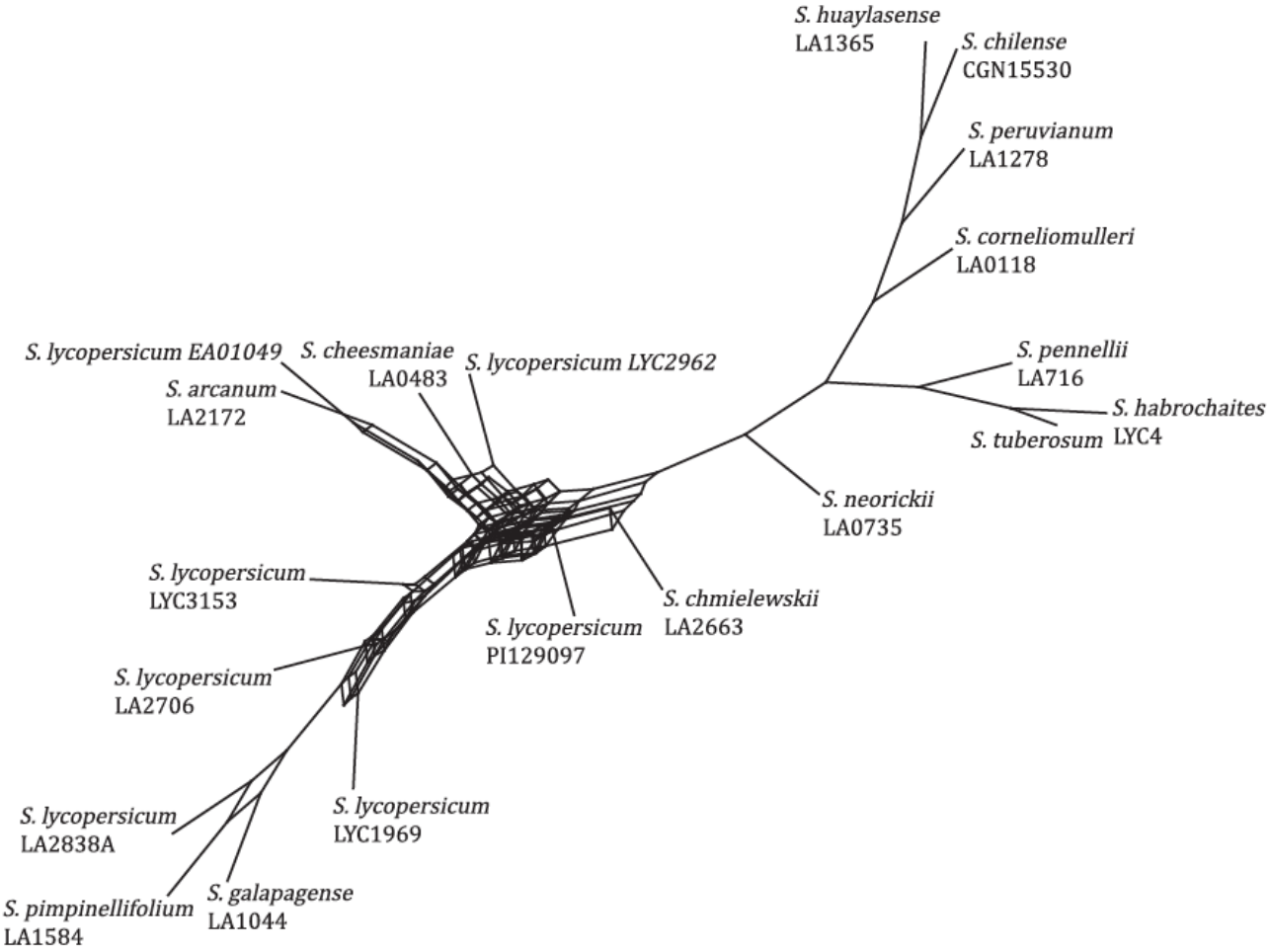


Figure 3.2 Relationships in *Solanum* section *Lycopersicon* shown as a filtered supernetwork. Splits present in 10% of all bootstrap trees are displayed. Conflict in the network, particularly within the 'Lycopersicon' group clade, suggests the occurrence of reticulation in the dataset and incongruence between genomic repeat clusters.

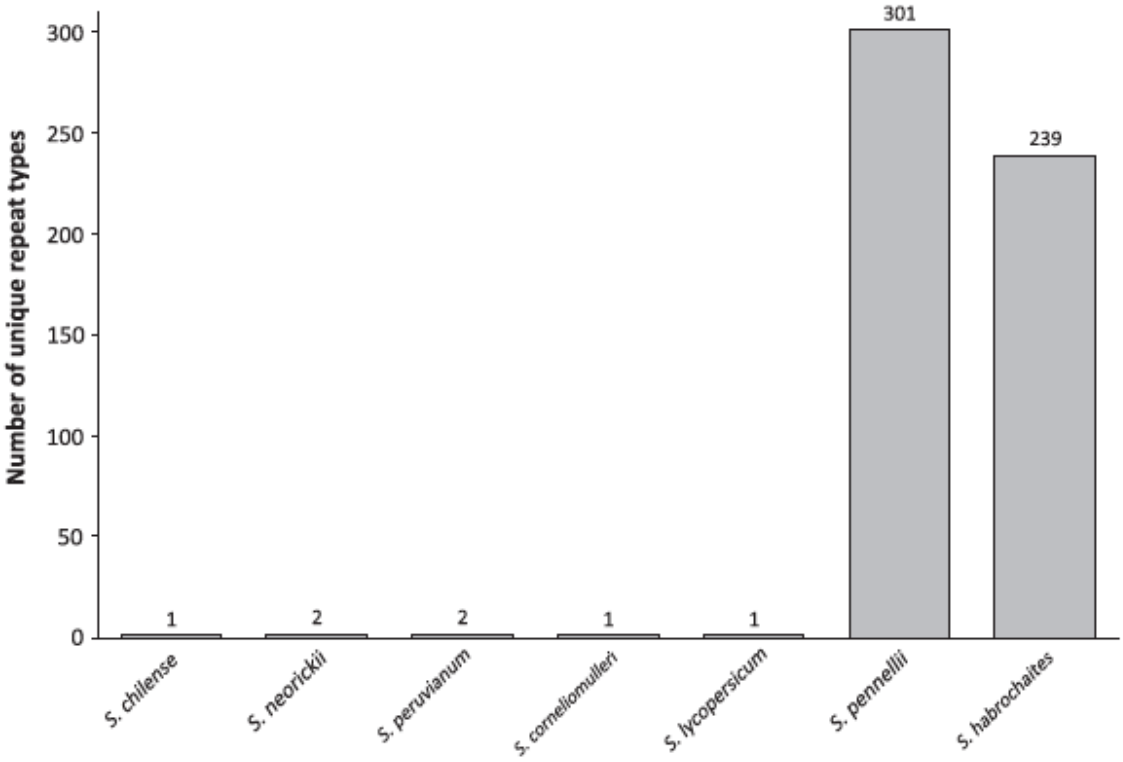


Figure 3.3 Number of unique repeat types (clusters) for the seven accessions that included them. Note *Solanum lycopersicum*, *Solanum corneliomulleri*, and *Solanum chilense* each have a single unique repeat type.

The latest rounds of genome sequencing are likely to add to the robust placement of some species (Aflitos *et al.*, 2014), and the current informal grouping within the section as defined by Peralta *et al*. (2008) does appear to reflect what is otherwise known about these taxa/accessions. All recent studies have broadly recovered the four informal groups as defined by Peralta *et al*. (2008): (1) 'Lycopersicon group' with *S. lycopersicum*, *S. cheesmaniae*, *S. galapagense*, and *S. pimpinellifolium*; (2) 'Arcanum group' with *S. arcanum*, *S. chmielewskii*, and *S. neorickii*; (3) 'Eriopersicon group' with *S. huaylasense*, *S. chilense*, *S. corneliomulleri*, *S. peruvianum*, and *S. habrochaites*; and (4) 'Neolycopersicon group' consisting of *S. pennellii*.

Rodriguez *et al*. (2009) also recovered these groups, although *S. huaylasense* was sister to the 'Arcanum group', rather than being a member of the 'Eriopersicon group', and *S. pennellii* and *S. habrochaites* were sister taxa. The groups of Peralta *et al*. (2008) were found to be clades based on genome-wide SNP data (Aflitos *et al.*, 2014), except that, similar to Rodriguez *et al*. (2009), they found *S. pennellii* and *S. habrochaites* to be sister taxa, thus restricting the concept of the 'Eriopersicon group' to *S. huaylasense*, *S. chilense*, *S. corneliomulleri*, and *S. peruvianum*. This could represent a loss of the anther appendage in *S. pennellii* or a parallel gain of the appendage in *S. habrochaites* and the rest of the core tomatoes. It is clear that further studies on the development of these characters are necessary to examine this result.

In the repeat analyses, most of these major groups were also identified. There were three notable differences: (1) *Solanum habrochaites* was recovered as sister to the rest of the section not as sister to *S. pennellii*; (2) two species of the 'Arcanum group', *S. chmielewskii*, and *S. arcanum*, were nested within the

'Lycopersicon group'; and (3) *Solanum neorickii* was recovered as sister to the 'Lycopersicon group' (including *S. arcanum* and *S. chmielewskii*).

The recovery here of *S. habrochaites* as sister to the rest of the core tomatoes differs from the results based on genome-wide SNPs (Aflitos *et al.*, 2014; Lin *et al.*, 2014), although it is perhaps not unexpected given the relatively unstable position of *S. habrochaites* and *S. pennellii* in previous analyses (Peralta *et al.*, 2008). It highlights the need for further developmental analysis of the sterile anther appendage long considered to be the synapomorphy of the core tomatoes (Peralta *et al.*, 2008).

The nesting of *S. chmielewskii* and *S. arcanum* within the 'Lycopersicon group' and the sister relationship of *S. neorickii* to this larger group are more unexpected results that require further investigation. The analyses of Aflitos *et al*. (2014) provided strong support for the 'Arcanum group', including *S. arcanum*, *S. chmielewskii*, and *S. neorickii* and for its sister relationship with the 'Eriopersicon group' (minus *S. habrochaites*), as reported by Rodriguez *et al*. (2009). The unusual placement of *S. arcanum* and *S. chmielewskii* in the phylogenetic analysis of repeats may be the result of the repetitive portion of the genome evolving under non-neutral processes, such as targeted repeat amplification/deletion or potentially horizontal gene transfer. Further characterization of repeat dynamics and additional taxon sampling could help to clarify this. The polytomy involving these taxa and the cultivated tomatoes could also be the result of extensive use of wild species in tomato breeding in the past, where gene regions from wild species have been introgressed in different cultivars of *S. lycopersicum* (Grandillo *et al.*, 2011). The filtered supernetwork based on 10% of all bootstrap trees (Fig. 3.2) shows clear evidence of potential reticulation and non-treelike evolution in this clade. Thus, the

placement of *S. arcanum* and *S. chmielewskii* could reflect the use of these wild species in previous tomato breeding.

The reference tomato genome ('Heinz 1706'; Tomato Genome Consortium, 2012) contains multiple introgressions from *S. pimpinellifolium*. Lin *et al*. (2014) found exotic fragments containing resistance genes in inbreeding lines, processing tomatoes, and fresh market hybrids that remained intact after several generations of backcrossing. This prospect of introgression, as with hybridization, would affect the analyses of genomic repeats, with some repeats specific to one parental lineage and some to the other parental lineage (or introgressed species). Network approaches do indeed provide some evidence for the involvement of introgression and/or hybridization in such scenarios, as indicated in the present study. However, this is complex and variable depending on the timeframe within which these processes occurred (e.g. polyploids of *Nicotiana*; Dodsworth *et al.*, 2015a).

Genome skimming for molecular systematics

The 'genome skimming' approach (*sensu* Straub *et al.*, 2012) involves low-coverage sequencing of genomic DNA using high-throughput technologies such as Illumina. The resulting data represents random sequences distributed throughout the genome but, because the coverage is low, the data will only represent the fraction of the genome that is in relatively high copy number. Notably, this includes ribosomal DNA from the nuclear genome (present in typically hundreds or thousands of copies) and organellar DNA (the plastid and mitochondrial genomes).

A current surge in using genome skimming approaches focuses on the plastid and mitochondrial sequences that can be assembled from low-coverage Illumina sequence data (Kane *et al.*, 2012; Steele *et al.*, 2012; Haran, Timmermans &

Vogler, 2013; Njuguna *et al.*, 2013; Gillett *et al.*, 2014) and this approach is proving successful in both animals and plants. It has advantages over other methods of high-throughput sequencing for phylogenetics because it requires no prior enrichment or complicated laboratory procedures; the downside is that it is currently limited by the cost of library preparation kits (unless custom protocols are developed). Reduced representation sequencing such as RADseq (Wagner *et al.*, 2013) and hybridization/pull-down methods (Guschanski *et al.*, 2013) both require extensive optimization and/or molecular laboratory work prior to the actual sequencing. A further advantage to genome skimming approaches is that they produce several datasets in one run: plastid, mitochondrial, and nuclear, which provide separate forms of evidence from 3 genomes that complement one another. In terms of nuclear markers, repetitive elements can be easily quantified using the RE pipeline and used in phylogeny reconstruction as shown in the present study. This provides additional evidence that may complement organellar and nuclear ribosomal cistron analyses.

Genome skimming, and in particular utilizing genomic repeats, may be useful for tapping into the genomic resources held in museum collections. Such DNA is often highly degraded, either simply because of age or a combination of age and the method by which specimens were initially dried. Collections have also been subject to various chemical treatments which impact upon DNA quality. These factors have previously hindered polymerase chain reaction success and still limit the availability of some high-throughput sequencing methodologies (such as amplicon sequencing or pull-down approaches). However, because genomic repeats are the most abundant sequences in genomic DNA samples, present in many copies, these will likely be adequately represented even in the most degraded of museum samples.

Conclusions: Future prospects for genomic repeats

In the *Solanum* example reported, *S. lycopersicum* samples formed a strongly supported group that included *S. cheesmaniae*, *S. galapagense*, and *S. pimpinellifolium* ('red/orange' fruited clade), as found in all previous studies (Peralta *et al.*, 2008; Aflitos *et al.*, 2014; Lin *et al.*, 2014), which indicates the utility of these data as phylogenetic markers at the intraspecific level. Despite this result, there were two unexpected placements within the *Lycopersicon* clade that require further investigation. Nonetheless, this is an important first result presenting the use of these data for low-level phylogenetic studies, such as phylogeography and investigations of widespread species and species complexes.

Genomic repeats could also serve as markers for DNA barcoding, although a crucial first step will be to determine whether there is a 'barcoding gap' (Meyer & Paulay, 2005; Meier, Zhang & Ali, 2008) in further datasets that include many samples of each species. Future developments including model-based inference in a custom Bayesian framework will add rigour to the analysis of these quantitative characters; this method can then be fully extended to some of the applications proposed in the present study at the intraspecific and interspecific levels.

Chapter 4 Phylogenomics of *Nicotiana* section *Suaveolentes* using genome skimming

Acknowledgements

The majority of flow cytometry measurements for *Nicotiana* section *Suaveolentes* taxa were conducted by Maïté Guignard.

Summary

Nicotiana section *Suaveolentes* currently represents approximately 26 species, most of which are endemic to Australia, with two species endemic to islands in the South Pacific and one species native to Namibia. Here I present phylogenomic results based on genome skimming, with complete taxon sampling and population-level sampling for several taxa. These represent the first phylogenetic results for the section that include all recognised taxa. The two species found exclusively in the South Pacific are not sister species, one representing an early-branching lineage with section *Suaveolentes* and the other a secondary dispersal as far as Tonga. The diversification of section *Suaveolentes* has occurred after the group arrived in Australia approximately 7 million years ago (mya), following a significant lag phase post-polyploidisation. Descending dysploidy is linked with the process of diploidisation in section *Suaveolentes*, which also includes significant genome downsizing. It is apparent that chromosome number continued to drop several times independently, ultimately from an ancestral n = 24 to as low as n = 15. A switch to annual life history strategy (with multiple reversals) seems to have played a role in adaptation and diversification in response to aridity, with most species diversity found in the central Australian deserts, the Eremaean Zone. There is little divergence in standard phylogenetic markers, and incongruence between organellar and nuclear DNA, with notable non-monophyly of population sequences in the plastome dataset. Coupled with the absence of clear morphological hybrids, this suggests evidence of the retention of ancestral polymorphism in plastid haplotypes. These data taken together paint a picture of section *Suaveolentes* as a recent and rapid radiation in Australia, still undergoing a putatively adaptive radiation in the Eremaean Zone.

Introduction

Nicotiana section *Suaveolentes* represents the largest section within the genus *Nicotiana*, containing approximately 26 Australasian and African species with a single allotetraploid origin in South America (Knapp, *et al.*, 2004; Marks, 2010a; Marks *et al*., 2011a; Ladiges *et al*., 2011). The genus *Nicotiana* is rife with hybridisation at both the homoploid and polyploid levels (Chase *et al.*, 2003; Leitch *et al.*, 2008; Kelly *et al.*, 2010), and allopolyploidisation has occurred over various timescales. This makes the genus an excellent model for studying the genomic and ecological effects of polyploidy in angiosperms (e.g. Renny-Byfield *et al.*, 2011; Renny-Byfield *et al.*, 2013; McCarthy *et al.*, 2015). Despite its size, section *Suaveolentes* is comparatively much less studied than the other polyploid sections. This section, by contrast to all of the other polyploid sections, has undergone significant diversification post-polyploidisation and thus can in itself be a model for angiosperm diversification following polyploidy and subsequent diploidisation. Section *Suaveolentes* species appear to have diversified within the last ~1-2 million years since a common origin for the section ~10 million years ago (Clarkson, 2007; Leitch *et al.*, 2008). Most of the species are endemic to Australia, with species diversity at its highest in the central Australian deserts (at least seven species occur in the immediate vicinity of Alice Springs).

Previous phylogenetic studies in *Nicotiana* have confirmed the monophyly of section *Suaveolentes* based on plastid and nuclear regions (Aoki and Ito, 2000; Chase *et al.*, 2003; Clarkson *et al.*, 2004; 2010; Kelly *et al.*, 2013), thereby likely rejecting Goodspeed's original hypotheses for at least three separate origins of section *Suaveolentes* taxa (Goodspeed, 1954). Goodspeed's hypotheses concerning the parental lineages involved in the origin of section *Suaveolentes* were perhaps disturbingly on point, indicating the involvement of *N. sylvestris* along with sections *Petunioides*, *Noctiflorae* and *Alatae*, that is consistent with

current genetic evidence (Kelly *et al.,* 2013; unpublished data). Despite this his necessary invocation of hybridisation in order to produce the various chromosome numbers in *Suaveolentes* is something that can be rejected based on the published phylogenetic studies. Prior phylogenetic results also show a general paucity of genetic variation in many of the standard phylogenetic markers – nrITS (Chase *et al.,* 2003), plastid genes *rbcL* and *matK* plus other non-coding spacer regions (Clarkson *et al.,* 2004), nuclear glutamine synthase (Clarkson *et al.,* 2010), *WAXY* and *MADS1* (Kelly *et al.,* 2013). This lack of variation is highly suggestive of a recent and ongoing diversification within *Nicotiana* section *Suaveolentes*.

Morphologically, one of the most useful characters for species identification has been shown to be corolla length, followed by other aspects of floral anatomy (Marks, 2010a; Marks *et al.,* 2011a), although to the casual observer floral morphology is similar amongst species (Figure 4.1). Vegetative differences are sometimes clearly marked between species, but I have found there is a great amount of plasticity that depends mostly on light levels and other environmental conditions. This therefore makes these characters less useful for identification purposes (e.g. the basal rosette character as defined by Marks (2010) is in fact very labile for many taxa in nature and the glasshouse).

In this chapter I aim to produce the most robust phylogenetic framework for *Nicotiana* section *Suaveolentes* to date, by utilising complete taxon sampling and a genome skimming approach, thereby vastly increasing both taxon and character sampling relative to previous studies. In addition to elucidating species relationships I will infer ancestral states for various characters that are related to the biology of these taxa.

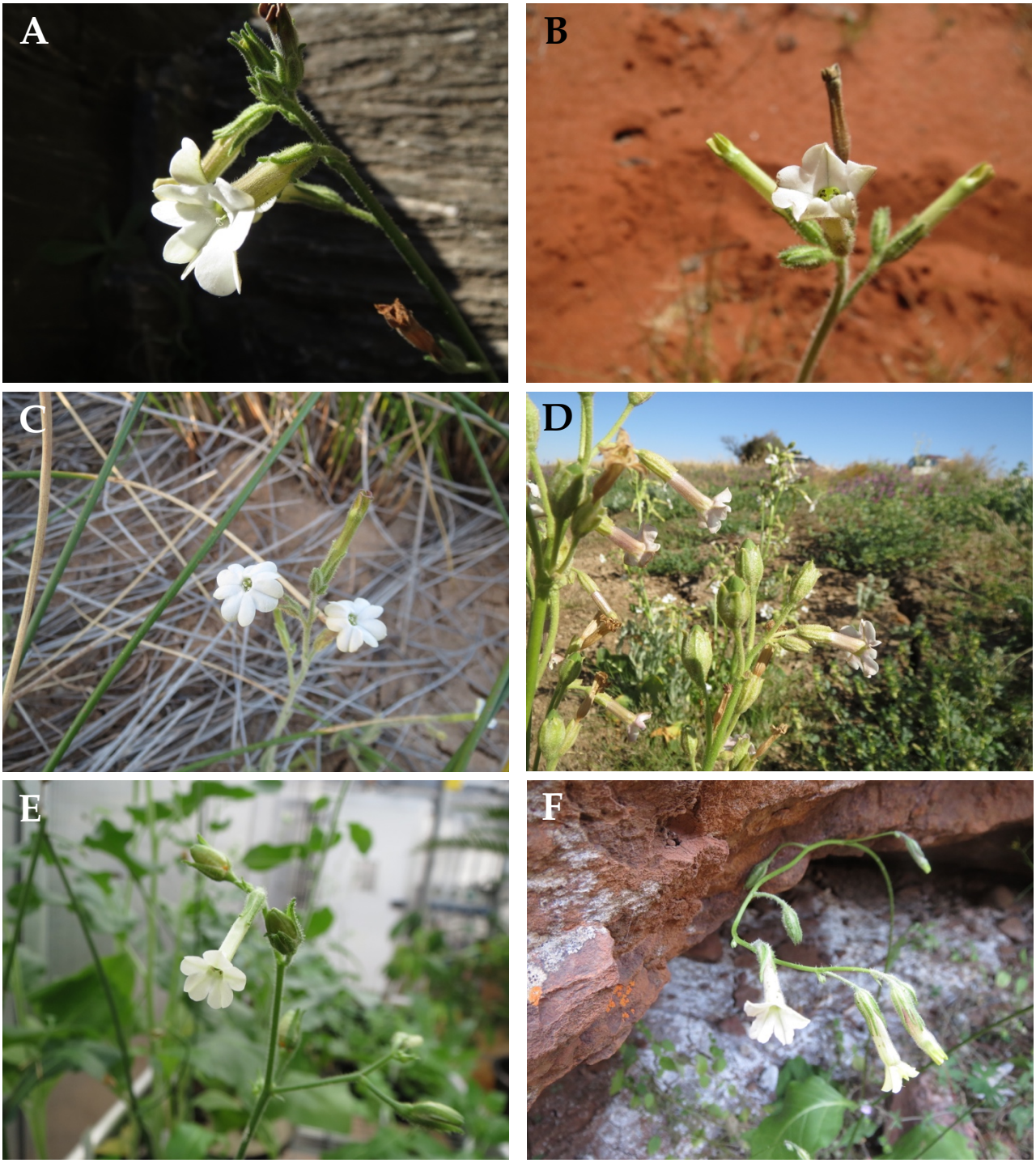


Figure 4.1 Examples of floral morphology in Australian species of *Nicotiana* section *Suaveolentes*. A – *N. maritima*; B-C – *N. velutina*; D – *N. truncata*; E – *N. forsteri*; F – *N. symonii*.

Materials and Methods

Plant materials

As far as was possible, wild-collected or wild-sourced plant material was used. Three sets of field collections were made over the course of 2013-15 in South Australia and Western Australia. Accessions were also provided by Steve Wylie (Murdoch University), stemming from Claire Marks' studies of morphology and chromosome numbers in *N.* section *Suaveolentes* (Marks, 2010a; Marks *et al*., 2011a; Marks *et al*., 2011b). These collections were included in genetic analyses as a priority if material was available. Additional material was received from seedbanks, botanical gardens and institutes including the Department of Genebank, Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung, Gatersleben (IPK); the Botanical and Experimental Garden, Radboud, University of Nijmegen, The Netherlands (Radboud); United States Department of Agriculture, North Carolina State University, NC (USDA); the Tropical Crops & Forages Collection of the Australian Plant Genetic Resource Information Service (AusPGRIS); the DNA Bank of the Royal Botanic Gardens, Kew (Kew); herbarium specimens from the Natural History Museum, London (BM). Further unidentified taxa were included in the study, which includes both potentially new species from field collections (*N. symonii*, *N. faucicola* etc.) and other taxa of unclear origin (*N. eastii*, *N. exigua*). Details of the number of accessions used and sources are given in Table 4.1.

Table 4.1 Summary of the number of accessions for each taxon including in *Nicotiana* section *Suaveolentes* and sources of plant material. CM = Claire Marks

Species	Accessions	Chromosome # (n)	Source(s)*
N. africana	1	23	USDA
N. amplexicaulis	2	18	AusPGRIS; CM
N. benthamiana	3	19	USDA; AusPGRIS
N. burbidgeae	4	21	Field; CM
N. cavicola	3	20, 23?	Field; CM
N. eastii	2	?	Radboud; IPK
N. excelsior	2	19	IPK; CM
N. exigua	3	?	Radboud; IPK
N. fatuhivensis	1	24	Kew
N. faucicola	4	?	Field
N. forsteri	3	24	AusPGRIS; CM
N. fragrans	1	?	BM
N. goodspeedii	6	16	Field
N. gossei	2	18	CM
N. heterantha	2	24	Field; CM
N. maritima	5	15	Field
N. megalosiphon subsp. megalosiphon	3	20	AusPGRIS; CM
N. megalosiphon subsp. sessilifolia	1	20	AusPGRIS
N. monoschizocarpa	1	24	CM
N. occidentalis subsp. hesperis	2	21	Field
N. occidentalis subsp. obliqua	3	21	Field
N. occidentalis subsp. occidentalis	3	21	Field
N. rosulata subsp. ingulba	2	20	Field; CM
N. rosulata subsp. rosulata	2	20	Field; AusPGRIS
N. rotundifolia	2	16	Field
N. simulans SA	11	20	Field
N. simulans WA	2	20	Field
N. sp. 68178	1	?	Field
N. sp. 68253	1	?	Field
N. sp. 68273	1	?	Field
N. suaveolens	1	15	CM
N. symonii	4	16	Field; CM
N. truncata	3	18	Field
N. umbratica	3	23	Field; Radboud; Kew
N. velutina	8	16	Field
N. wuttkei	1	16	Radboud

*For full names see Methods. Voucher specimens are held at K, AD, and Queen Mary University of London.

Dating the origin of section Suaveolentes

A Bayesian approach was used to infer the date of divergence of the allopolyploid section *Suaveolentes*. A combined plastid data matrix (*trnL-F*, *trnS-G*, *matK*, *ndhF*) for diploid and polyploid *Nicotiana* was first reconstructed from the Kelly *et al*. (2013) plastid matrix, based on data from Clarkson *et al*. (2004), with outgroup taxa from Anthocercidae. This simplification of available data was chosen to avoid the complications of multiple copies in low-copy nuclear genes for polyploids, plus the known conversion of nrITS to (often) the paternal parental lineage, which introduces obvious discordance with the plastid data (and this topological incongruence prevents the direct use of dates from Särkinen *et al.*, 2013).

Time-calibrated phylogenetic trees were reconstructed using BEAST2 (Bouckaert *et al.*, 2014) on the CIPRES web server (Miller *et al.*, 2010), using an uncorrelated lognormal relaxed clock model (Drummond *et al.*, 2006; Heled and Drummond, 2012). A secondary calibration was used to date the split between *Symonanthus* and *Nicotiana* at ~15 Mya, following a study of all Solanaceae (1000 tips) that used carefully re-evaluated fossil data (Särkinen *et al.*, 2013). A normal distribution for the node age prior was used (mean of 15, standard deviation of 1.5), thereby encompassing the values and error associated with the original dating (95% HPD interval of ~11-20 MYA). Analyses were run for 10,000,000 generations, storing every 1000 generations.

Genome size estimation by flow cytometry

Genome sizes (1C-values) were estimated by flow cytometry, as described in Pellicer & Leitch (2014), using a Partec CyFlow Space fitted with a Cobalt Samba green (532 nm, 100 mW) laser. Approximately 20 mg of each leaf sample was finely chopped using a razor blade along with the internal standard in Galbraith buffer. The standard used was parsley, *Petroselinum crispum* "Champion Moss

Curled" (1C=2.22 pg). For each species, full estimates were made with leaves from three individual plants of the same accession, where possible (i.e. each sample was run three times, measuring the C-value of 5000 (combined) nuclei per run). To screen for intraspecific variation, tissue from at least one individual from every lab-grown accession/population was co-chopped with the internal standard and run on the flow cytometer to measure 1000 or more nuclei. See Table 4.2 for a summary of C-values and coefficients of variation for each species.

Genomic DNA extraction and gDNA quality control

Approximately 100 mg of leaf tissue was used to extract genomic DNA (gDNA), using mostly the modified CTAB protocol of Wang *et al*. (2012), but in some cases using a Qiagen DNEasy Plant Mini Kit (Qiagen, Santa Clarita, CA). Fresh or silica-dried leaf tissue were used for the bulk of gDNA extractions. Herbarium extractions were performed with a DNEasy Plant Mini Kit (Qiagen) as this has been shown to be one of the most effective methods for extractions from herbarium specimens (Särkinen *et al.*, 2012; Staats *et al*., 2013). In all cases, leaf samples were first frozen in liquid nitrogen and ground using either a pestle and mortar or a Qiagen TissueLyzer (Qiagen) with one or two 3 mm steel beads for sample homogenisation. Genomic DNA extractions were analysed on a 1% agarose gel and subsequently fluorometrically with a Qubit analyser (Life Technologies Ltd, UK) to assess both quality (integrity) and concentration. Samples that had low concentrations were either re-extracted, concentrated using a speed-vac or used in subsequent whole-genome amplification using a Qiagen Repli-G Mini Kit (Qiagen). Some samples, particularly from herbarium specimens, were run on an Agilent TapeStation (Agilent Technologies, Santa Clara, CA) with a Genomic DNA Analysis ScreenTape Assay, in order to assess fragment size distribution of the gDNA.

High-throughput sequencing – genome skimming

Samples were multiplexed and sequenced in several runs during 2014 and 2015; the bulk of samples (88) were multiplexed and run on a single high-output NextSeq run as a set of 96 samples (V2 chemistry – 300 cycles; 151 bp paired-end reads). Remaining samples were pooled in batches of 8-12 and run on a MiSeq in three separate runs (V2 chemistry – 300 cycles; 151 bp paired-end reads). For the NextSeq samples a high-throughput TruSeq PCR-free library preparation kit (Illumina) was used, with 350 bp Covaris fragmentation. For the MiSeq samples low-throughput TruSeq PCR-free library preparation kits were used, with 550 bp Covaris fragmentation. The difference in average insert size of the libraries was due to the suboptimal performance of the NextSeq machine with genomic libraries over 350 bp. Covaris fragmentation was performed with a Covaris M220 Focused-ultrasonicator with microTUBE Snap-Cap AFA Fiber tubes. gDNA from fresh and silica-dried material was subject to Covaris fragmentation as described in the TruSeq protocol; herbarium gDNA was only fragmented in the case of large fragments being present in high concentration, and the sonication time was reduced appropriately. Prepared libraries were checked for success in a qualitative manner using a Bioanalzyer with a High Sensitivity DNA Analysis Kit (Agilent). Library concentration was then checked using Qubit and subsequently with qPCR, as the prepared libraries contain gDNA fragments that do not have adapters properly ligated that will be quantified with Qubit but non-sequencable; hence qPCR is the only reliable way to measure sequencable fragments. qPCR was performed using a NEBNext Library Quant Kit for Illumina (New England Biolabs, UK), running libraries in triplicate at 10,000 and 20,000 dilutions as per the Illumina protocol. Equal amounts of libraries were then pooled to get sets of 96 (dual-indexed), 8 or 12 (single-indexed) pools.

Assembly of high-copy genomic regions

Whole plastome sequences were assembled for each taxon/sample using MIRA Version 4.0.2 (Chevreux *et al.*, 1999) by mapping reads in a reference-guided assembly to the *N. tabacum* plastid genome sequence (GenBank: NC_001879.2). This included coding and non-coding regions and the inverted repeat regions. Contigs were assessed visually by converting .caf files to .ace files and viewed in Tablet v. 1.15.09.01 (Milne *et al.*, 2013) to check for chimeric or misassembled regions. Other assemblers were tested, *de novo* and reference-guided but found to be inferior as they did not map large regions including the inverted repeats. Additionally, ribosomal DNA cistron sequences were assembled using Geneious v. 8.1.7 with a consensus reference from the alignment of *N. tabacum* 26S (AF479172.1) and *N. benthamiana* ETS-18S-5.8S-26S-ETS (KP824745.1) rDNA sequences. Reads from each sample were mapped to this reference in a batch assembly, using sensitivity set to Medium, a minimum mapping quality of 20 and Fine tuning set to 'up to 5 iterations'. Assemblies were individually checked for read-depth and any other issues.

Clustering of high-throughput reads and repeat abundance estimation

Reads were quality filtered to include those above quality score 20 over 95% of the read length using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). In addition, base composition across the read length was assessed to check for any adapters post-filtering. Reads from each sample/taxon were prefixed with a unique 8-letter code. Each set of reads was then down-sampled to 10,000 reads (assuming the same genome size for each species) and combined into one fasta file for input to the RepeatExplorer pipeline. This initial combined dataset consisted of 109 taxa and 1080909 reads (one sample had fewer reads, 9091).

A second dataset was created using a subset of taxa (37) with a greater number of reads (0.5% proportion of each genome). Approximately one sample per species, chosen based on clades recovered in other phylogenetic analyses and the number of taxa in the species tree analyses, and with the number of reads proportional to genome size (where 1C-value data was unavailable, an assumed size of 1C = 3.2 Gb was used as this represents the mean for known taxa). This dataset consisted of 37 taxa and 4,264,969 reads.

Each dataset was run through the RepeatExplorer pipeline (on both the public webserver at http://www.repeatexplorer.org and QMUL's own Galaxy server at http://galaxy.hpc.qmul.ac.uk, to check for reproducibility and account for computational demands) using default settings of 95% similarity over 55% of the read length (80 bp), read renaming but retaining taxon-specific 8-letter codes.

Phylogenomic analyses

Plastomes and rDNA

The plastome dataset was aligned using MAFFT version 7 (Katoh and Standley, 2013) with default settings (i.e. 'Auto' strategy) and manually curated by eye in Geneious v. 8.1.7 to discard poorly aligned or ambiguous regions due to low coverage in some taxa. All gaps and ambiguities were coded as missing data. The final plastome alignment consisted of 107 taxa and 156,839 bp with 1.4% missing data; the final alignment for rDNA consisted of 116 taxa and 7,207 bp with 2.8% missing data.

Maximum likelihood (ML) analyses were run using RAxML v. 8.2.4 (Stamatakis, 2014) on the CIPRES Science Gateway (Miller *et al.*, 2010). The GTR+GAMMA model was used with 4 discrete rate categories to approximate the gamma distribution; 1000 bootstrap replicates were performed using the rapid

bootstrapping algorithm. Bipartition frequencies were printed onto the best-scoring ML tree.

Bayesian analyses were conducted using PhyloBayes MPI v. 1.5a (Lartillot *et al.*, 2009) on the CIPRES server (Miller *et al.*, 2010). The CAT model was used for the profile mixture, which is an infinite mixture model that accounts for site-specific base equilibria as each site is assigned a vector profile over the 4 bases according to a Dirichlet process (Lartillot and Philippe, 2004). These are then combined with a global set of exchange rates defined by a GTR model with 8 distinct rate categories for discretisation of the gamma distribution (CAT-GTR settings), thereby yielding site-specific substitution processes. This method was chosen as it is the most complex of current models, effectively partitioning the data per site, and has been shown to account well for compositional heterogeneity, homoplasy and long-branch attraction artefacts. Two independent runs were performed, each with 2 chains, and run until convergence was achieved (represented by maxdiff of <0.1 between chains). A burn-in of 1000 cycles was used, and then trees subsampled every 10 in order to build the posterior majority-rule consensus tree. For the plastome data, each run consisted of 17,268 and 24,302 cycles; for the rDNA data each run consisted of 6677 and 6645 cycles.

Genomic repeats

Clusters from the RepeatExplorer output were initially filtered to remove ones that contained plastid reads or Illumina control sequences (e.g. PhiX control DNA). Post-filtering the top 1000 most abundant clusters were used downstream for phylogenetic analyses. A square root transform of the raw abundances was performed in order to make the abundances in the range 0-65 (required as input for TNT). Previously factorial transformations were performed (Dodsworth *et al.*, 2015a; 2015b), but these were noted to have the effect of overweighting clusters with large relative differences compared to

those with smaller differences between taxa. This is due to the additive nature of the parsimony algorithms, and thus transformations that retain more of the additive difference between cluster abundances are preferred. TNT was then used to reconstruct the phylogenetic tree using maximum parsimony and the repeat cluster abundances used as continuously varying characters for tree reconstruction in TNT (Goloboff *et al*., 2006; 2008). Trees were inferred broadly as described in Dodsworth *et al*. (2015a), except that due to the number of taxa (109 for dataset 1 and 37 for dataset 2) implicit enumeration was not used and instead a heuristic search was performed using a new technology sectorial search with 3 rounds of tree fusing. Additionally, a further subset of 15 taxa was created to check the heuristic parameters relative to an implicit (exact) search. Resampling was performed using 100 symmetric bootstrap replicates and bootstrap percentages were displayed on the single most parsimonious topology.

Species tree inference

In an attempt to reconstruct the species tree for section *Suaveolentes* and summarise the information in the various genomic datasets (and their variable evolutionary histories) two methods of species tree estimation were used. For both analyses obviously distinct taxa were coded as separate species (e.g. *N. simulans* from WA vs. SA; unidentified *N.* spp. from WA) in order to minimise the discordance represented by clearly distinct taxa unnecessarily named the same. Firstly, a more simplistic parsimony analysis was conducted in Mesquite v. 3.04, using the maximum clade credibility gene trees for rDNA and plastomes (from the *BEAST analysis below) as input trees. A heuristic search was performed to find the species tree that minimises deep coalescences for multiple loci (Maddison, 1997). Parameters were set to auto-resolution of polytomies, using branch lengths of contained trees, and SPR rearrangement with MaxTrees set to 100. Deep coalescences are summed for all gene trees thereby assuming that each locus is independent. This method assumes that all gene tree

discordance is due to incomplete lineage sorting (i.e. deep coalescence, where gene trees coalesce more deeply than the speciation events).

A second species tree analysis was conducted using a Bayesian species tree methodology (*BEAST) utilising the multispecies coalescent model in BEAST2 (Bouckaert *et al.*, 2014). This method estimates the most probable species tree from multi-locus multi-individual data sequence data, by simultaneously co-estimating the gene trees in a species tree (Heled and Drummond, 2010). *BEAST was run with two partitions (rDNA and plastomes) using the alignments from individual analyses, with a GTR+GAMMA substitution model with 4 rate categories for the discretisation of the gamma distribution. A Yule process prior was specified for the species tree prior and piecewise linear with constant root prior for the population size model. Ploidy for the plastome data was set to 'mitochondrial', as this mode of inheritance is taken into account for the effective population size. A strict clock model was used, setting the plastome rate to 1.0 and estimating the rDNA substitution rate relative to the plastome.

Poor mixing and convergence was initially observed with a particularly low effective sample size (ESS) of the clock rate parameter and associated substitution rate parameters, and a clear upwards trend in the trace. This is probably due to low signal of timing information in the combined data, and therefore the extremely large upper bound (infinity) of the uniform prior dominates the rate estimate resulting in nonsensical estimates. As such the upper bound for the clock rate uniform prior distribution was set to 1.01 instead of infinity (with initial value 1), which resulted in much better ESS values and mixing. Five independent chains were run, each for between 20-25 million generations (sampling every 5000th generation), resulting in a total of 99.29 million generations when traces were combined with LogCombiner after confirming convergence using Tracer and using a burn-in of 10%. Trees were

combined from all 5 runs using LogCombiner and a burn-in of 10% from each, resulting in 19,862 trees. TreeAnnotator was then used to produce a maximum clade credibility consensus tree from the 19,862 species trees with median node heights, which finds the tree that maximises the product of clade posterior probabilities in the post burn-in trees. An alternative visualisation of the species tree was produced using DensiTree v. 2.2.2 (Bouckaert and Heled, 2014) using the 19,862 post burn-in trees from *BEAST analysis and setting the shuffle option to 'closest outside first' in order to rearrange tips to create the clearest picture of the entire tree set.

Genealogical sorting index (GSI) analysis

The genealogical sorting index (GSI) provides a measure of the relative exclusivity (i.e. exclusive ancestry) of a group of sequences on a phylogenetic tree, whether they are population or individual samples (Cummings *et al.*, 2008). The maximum value of GSI is 1, which indicates monophyly, whereas a value of 0 indicates dispersion over the entire tree topology. Topology is used to quantify the coalescent events uniting a group and exclusivity of groups is continuously distributed. Significance is tested statistically by permuting trees with stochastically rearranged terminals and computing the proportion that exceed the GSI in the original tree to provide a *p*-value for rejection of the null hypothesis (i.e. that the group is of mixed ancestry; the probability of observing a GSI value by chance alone that is equal to or exceeds the original GSI). The GSI takes into account topological uncertainty and is a useful metric for assessing species boundaries, cryptic species and testing for monophyly of populations (e.g., Cranston *et al.*, 2009; Sakalidis *et al.*, 2011; Schmidt-Lebuhn *et al.*, 2012). GSI values were computed using GSI version 0.92 on the GSI web service (www.molecularevolution.org) through the Lattice Project (Bazinet and Cummings, 2008). Permutation tests were run with 10,000 replicates, and GSI values calculated for both plastomes and rDNA, and together (ensemble statistics) by integrating the GSI values across an ensemble of topologies.

Ancestral state reconstructions

The ancestral states for chromosome number, genome size, corolla length, and life history strategy were reconstructed using the phylogenomic framework generated. Ancestral state reconstructions for life history strategies were conducted in Mesquite v. 3.04 using parsimony-based reconstruction, as the small number of changes and binary states meant more complicated model-based methods were unwarranted.

Genome size and corolla length were reconstructed using the program BayesTraits v. 2.0 (http://www.evolution.reading.ac.uk/BayesTraits.html; Pagel, 1999). In order to account for both phylogenetic uncertainty and uncertainty in the model parameters, the post burn-in trees from the *BEAST analysis were used as input (19,862 trees). The trees were converted with BayesTrees to ensure consistent rooting with *N. noctiflora* (outgroup). Corolla tube lengths were taken from relevant descriptions for each species, using the middle value where a range was given (most cases). For corolla length and genome size, continuous character reconstruction was implemented in BayesTraits with the continuous random walk model.

Values were reconstructed at particular nodes that were well supported and represented the root of *Suaveolentes* and the origin of well-supported major subclades. Means for reconstructed nodes and their standard deviations were plotted onto the DensiTree species tree.

Chromosome number was reconstructed both with parsimony in Mesquite v. 3.04 and maximum likelihood using chromEvol version 2.0 (Glick and Mayrose, 2014). The CONST_RATE model was selected with the lowest AIC value amongst 10 models tested, which includes parameters for rate of single chromosome increase, this rate being dependent on the current chromosome

number, and a rate of polyploidisation. The outgroup was pruned, and the root node was fixed to n = 24, the maximum chromosome number set to 10 times the maximum found in the data. Taxa without chromosome counts were defined as missing data. *Nicotiana cavicola* was coded as n = 23 and n = 20 (0.5 probability of each count, owing to uncertainty – Marks, 2010a).

Results

Genome size

Flow cytometry estimates of genome size (1C-values) for *Nicotiana* section *Suaveolentes* taxa are presented in Table 4.2. Most species were found to have a genome size around 3.0-3.5 pg, with the mean genome size for *Suaveolentes* at 3.7 pg. *Nicotiana forsteri* and *N. africana* were found to have larger genome sizes, 4.9 pg and 5.3 pg, respectively. *Nicotiana eastii* had a large genome size of 6.6 pg.

Table 4.2 Genome sizes (1C-values) for *Nicotiana* section *Suaveolentes* taxa

Species	Mean C-value (pg)	Mean c.v.
N. africana	5.28	2.06
N. amplexicaulis	3.58	2.64
N. benthamiana	3.53	3.47
N. burbidgeae	3.17	2.92
N. cavicola	2.72	2.88
N. eastii	6.60	3.20
N. excelsior	3.44	3.53
N. exigua	3.48	3.29
N. faucicola	3.58	3.60
N. forsteri	4.86	2.19
N. goodspeedii	3.47	2.87
N. gossei	3.76	3.44
N. maritima	3.41	2.42
N. megalosiphon subsp. *megalosiphon*	3.30	3.22
N. megalosiphon subsp. *sessilifolia*	3.73	3.10
N. occidentalis subsp. *obliqua*	3.17	4.07
N. rotundifolia	2.68	2.65
N. simulans SA	2.92	3.13
N. symonii	3.52	3.21
N. truncata	3.61	2.20
N. umbratica	3.83	1.92
N. velutina	3.18	3.46
N. wuttkei	3.42	1.69

Dating of section Suaveolentes

The phylogenetic analysis with calibration based on the Särkinen *et al*. (2013) date for the split between *Symonanthus* and *Nicotiana* resulted in a well-resolved tree (Figure 4.2) for *Nicotiana*. The date for the split of section *Suaveolentes* from section *Noctiflorae* is ~6.75 mya (95% HPD of 4.40-9.19). *N. africana* diverged from the rest of section *Suaveolentes* ~6 mya, and *N. forsteri* is the next diverging

lineage splitting from the remaining species of *Suaveolentes* ~5 mya. The remaining speciation events within the section are dated from ~0.2-2.0 mya, with the majority around 2 mya.

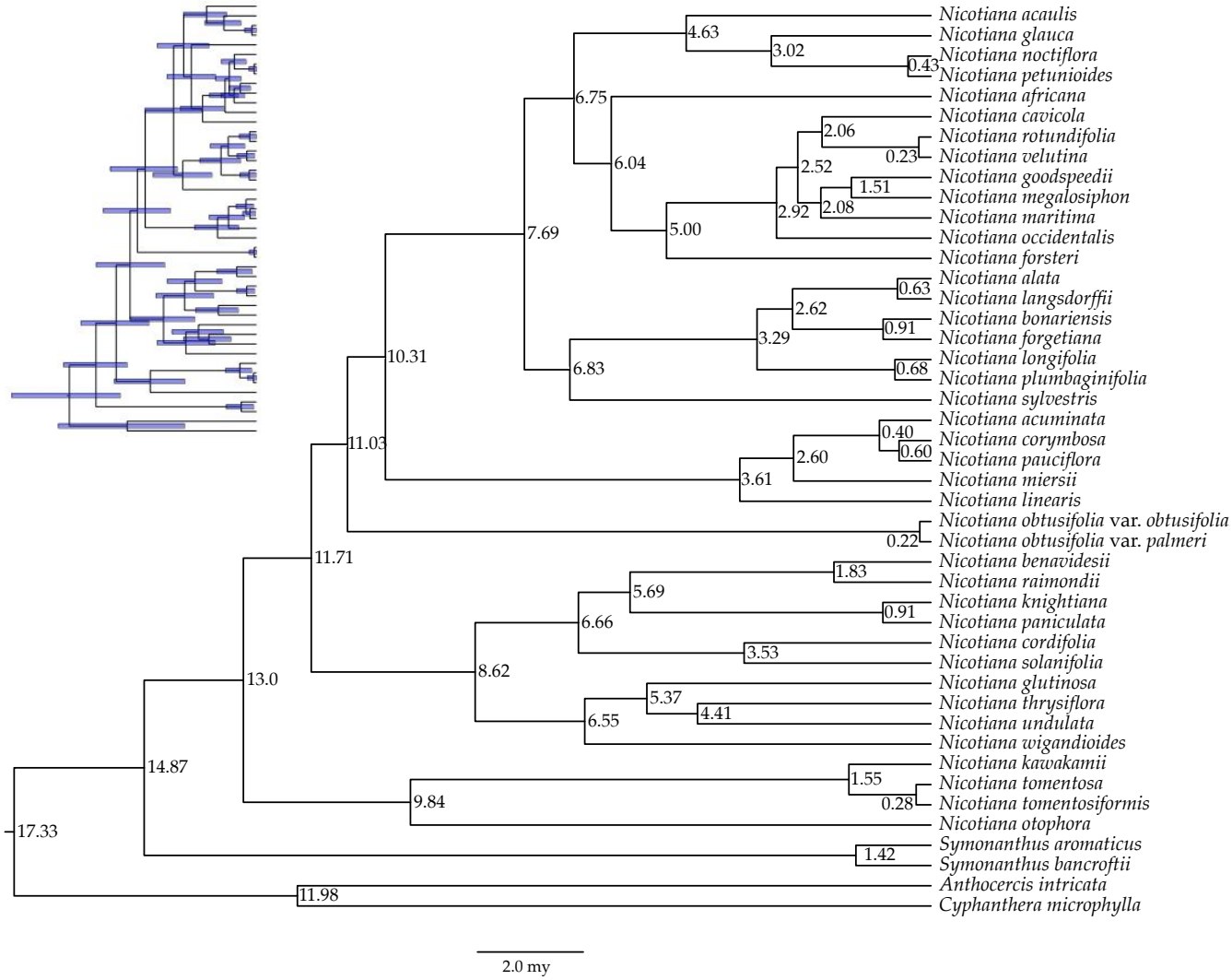


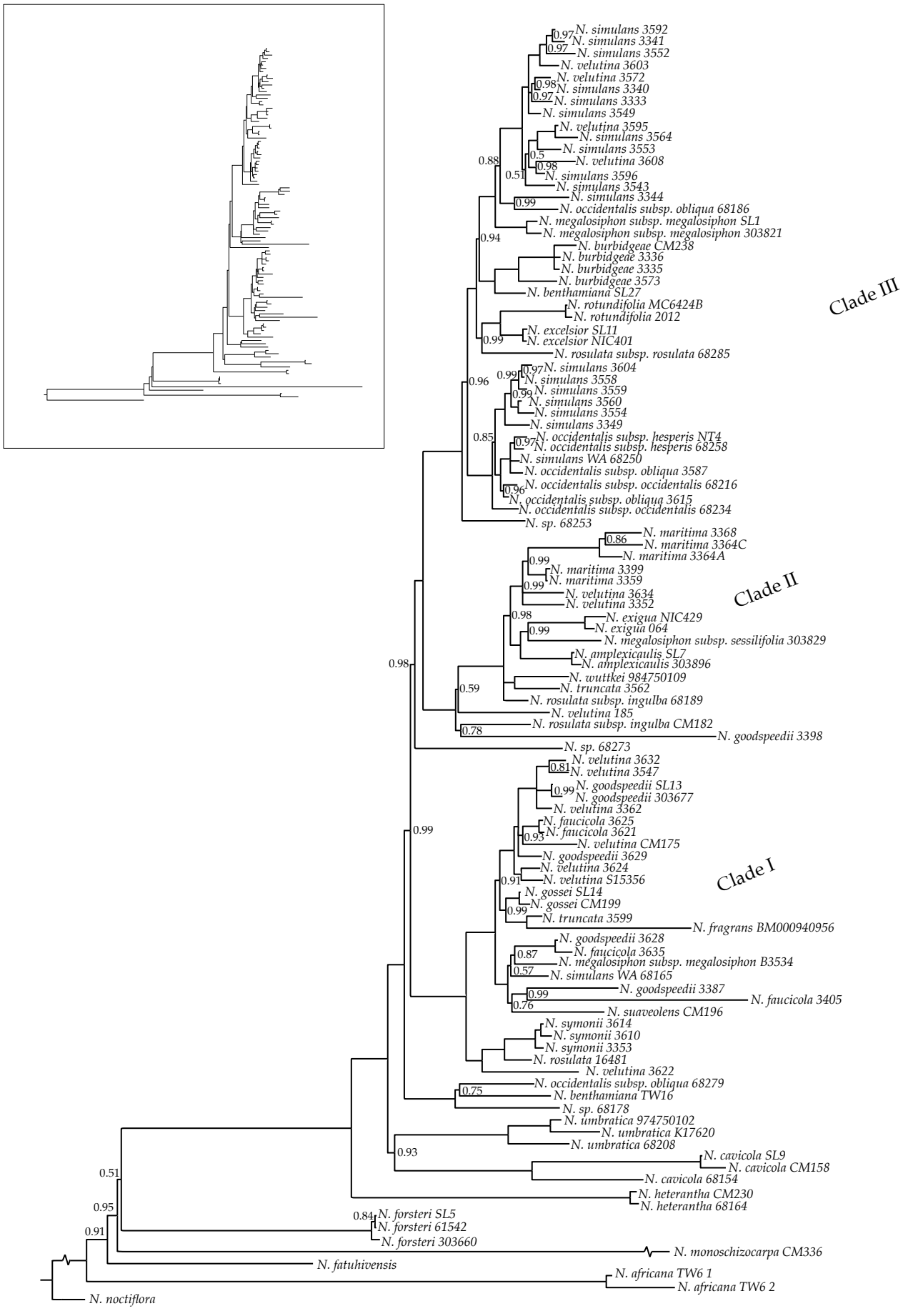
Figure 4.2 Time-calibrated tree of *Nicotiana* based on the relaxed lognormal clock in BEAST, with secondary calibration of the split between *Symonanthus* and *Nicotiana*. Numbers at nodes represent calibrated node ages; see inset for 95% HPD intervals. Maximum clade credibility tree based on 9,000 trees post burn-in.

Plastome tree

Phylogenomic analyses of whole plastomes resulted in well-supported topologies in both ML and Bayesian analyses, with the Bayesian analysis shown in Figure 4.3. With rooting on *N. noctiflora* as the outgroup, the sister to the rest of section *Suaveolentes* is *N. africana* – the sole African species of the section (and indeed *Nicotiana*). The next diverging clade consists of *N. fatuhivensis* (0.91 PP), which is an isolated endemic to the Marquesas Islands in the South Pacific, followed by *N. monoschizocarpa* on a particularly long branch. *Nicotiana monoschizocarpa* is a narrow endemic found from near Darwin, Northern Territory. The next split consists of *N. forsteri* accessions, which form a strongly supported clade in themselves but weakly supported (0.51 PP) as the next diverging lineage sister to the rest of the section. A long branch then splits *N. forsteri* from the rest of section *Suaveolentes*. *Nicotiana heterantha* is then the first species in the main large clade representing all the Australian taxa (bar *N. forsteri*), sister to the rest of this clade (Figure 4.3). Two other Western Australian species, *N. cavicola* and *N. umbratica* together form the next clade, sister to the remaining species. The remaining taxa are then largely grouped into three major clades (Figure 4.3), with a few exceptions. Clade I consists mainly of *N. velutina*, *N. goodspeedii* and *N. faucicola* accessions – these are seemingly all mixed up with regard to populations/accessions of each species not forming unique clusters in the tree. This clade also includes the Australian *N. gossei*, *N. suaveolens*, *N. symonii*, *N. rosulata* subsp. *rosulata*, *N. truncata* and the South Pacific species *N. fragrans*.

Clade II consists of a mix of Western and South Australian species, with *N. rosulata* subsp. *ingulba*, *N. amplexicaulis*, *N. megalosiphon* subsp. *sessilifolia*, *N. exigua*, *N. wuttkei*, and the southern coastal *N. maritima*. There are also some accessions of *N. velutina* in this clade, one *N. truncata*, and one *N. goodspeedii*.

Clade III also consists of western, northern, and southern Australian taxa. A potentially new species *N.* sp. 68253 is sister to the rest of the clade, the next clade then consists of *N. occidentalis* and *N. simulans* accessions, which is then sister to the rest of clade III. This last clade is comprised of two subclades, the first consisting of all *N. occidentalis* accessions (this includes all three subspecies, *N. occidentalis* subsp. *occidentalis*, *N. occidentalis* subsp. *hesperis* and *N. occidentalis* subsp. *obliqua*) with no clear grouping of subspecies as exclusive entities. The other subclade consists of a group of *N. simulans* accessions from South Australia. A different accession (from that in clade I) of *N. rosulata* subsp. *rosulata* is then found as sister to a clade comprising *N. excelsior* and *N. rotundifolia* as sister taxa. One accession of *N. benthamiana* (SL27) is then found as sister to the narrow endemic *N. burbidgeae* from Dalhousie Springs, South Australia (for which all accessions form an exclusive group). *Nicotiana megalosiphon* subsp. *megalosiphon* is then sister to most of *N. simulans* accessions from South Australia. A few *N. velutina* accessions are included in this clade with *N. simulans*, but these likely represent misidentifications (due to collection of dried and dead plants, from which plants were later grown).


N. simulans 3592
N. simulans 3341
N. simulans 3552
N. velutina 3603
N. velutina 3572
N. simulans 3340
N. simulans 3333
N. simulans 3549
N. velutina 3595
N. simulans 3564
N. simulans 3553
N. velutina 3608
N. simulans 3596
N. simulans 3543
N. simulans 3344
N. occidentalis subsp. obliqua 68186
N. megalosiphon subsp. megalosiphon SL1
N. megalosiphon subsp. megalosiphon 303821
N. burbidgeae CM238
N. burbidgeae 3336
N. burbidgeae 3335
N. burbidgeae 3573
N. benthamiana SL27
N. rotundifolia MC6424B
N. rotundifolia 2012
N. excelsior SL11
N. excelsior NIC401
N. rosulata subsp. rosulata 68285
N. simulans 3604
N. simulans 3558
N. simulans 3559
N. simulans 3560
N. simulans 3554
N. simulans 3349
N. occidentalis subsp. hesperis NT4
N. occidentalis subsp. hesperis 68258
N. simulans WA 68250
N. occidentalis subsp. obliqua 3587
N. occidentalis subsp. occidentalis 68216
N. occidentalis subsp. obliqua 3615
N. occidentalis subsp. occidentalis 68234
N. sp. 68253
N. maritima 3368
N. maritima 3364C
N. maritima 3364A
N. maritima 3399
N. maritima 3359
N. velutina 3634
N. velutina 3352
N. exigua NIC429
N. exigua 064
N. megalosiphon subsp. sessilifolia 303829
N. amplexicaulis SL7
N. amplexicaulis 303896
N. wuttkei 984750109
N. truncata 3562
N. rosulata subsp. ingulba 68189
N. velutina 185
N. rosulata subsp. ingulba CM182
N. goodspeedii 3398
N. sp. 68273
N. velutina 3632
N. velutina 3547
N. goodspeedii SL13
N. goodspeedii 303677
N. velutina 3362
N. faucicola 3625
N. faucicola 3621
N. velutina CM175
N. goodspeedii 3629
N. velutina 3624
N. velutina S15356
N. gossei SL14
N. gossei CM199
N. truncata 3599
N. fragrans BM000940956
N. goodspeedii 3628
N. faucicola 3635
N. megalosiphon subsp. megalosiphon B3534
N. simulans WA 68165
N. goodspeedii 3387
N. faucicola 3405
N. suaveolens CM196
N. symonii 3614
N. symonii 3610
N. symonii 3353
N. rosulata 16481
N. velutina 3622
N. occidentalis subsp. obliqua 68279
N. benthamiana TW16
N. sp. 68178
N. umbratica 974750102
N. umbratica K17620
N. umbratica 68208
N. cavicola SL9
N. cavicola CM158
N. cavicola 68154
N. heterantha CM230
N. heterantha 68164
N. forsteri SL5
N. forsteri 61542
N. forsteri 303660
N. monoschizocarpa CM336
N. fatuhivensis
N. africana TW6 1
N. africana TW6 2
N. noctiflora
Clade III
Clade II
Clade I
0.03 subtitutions/site

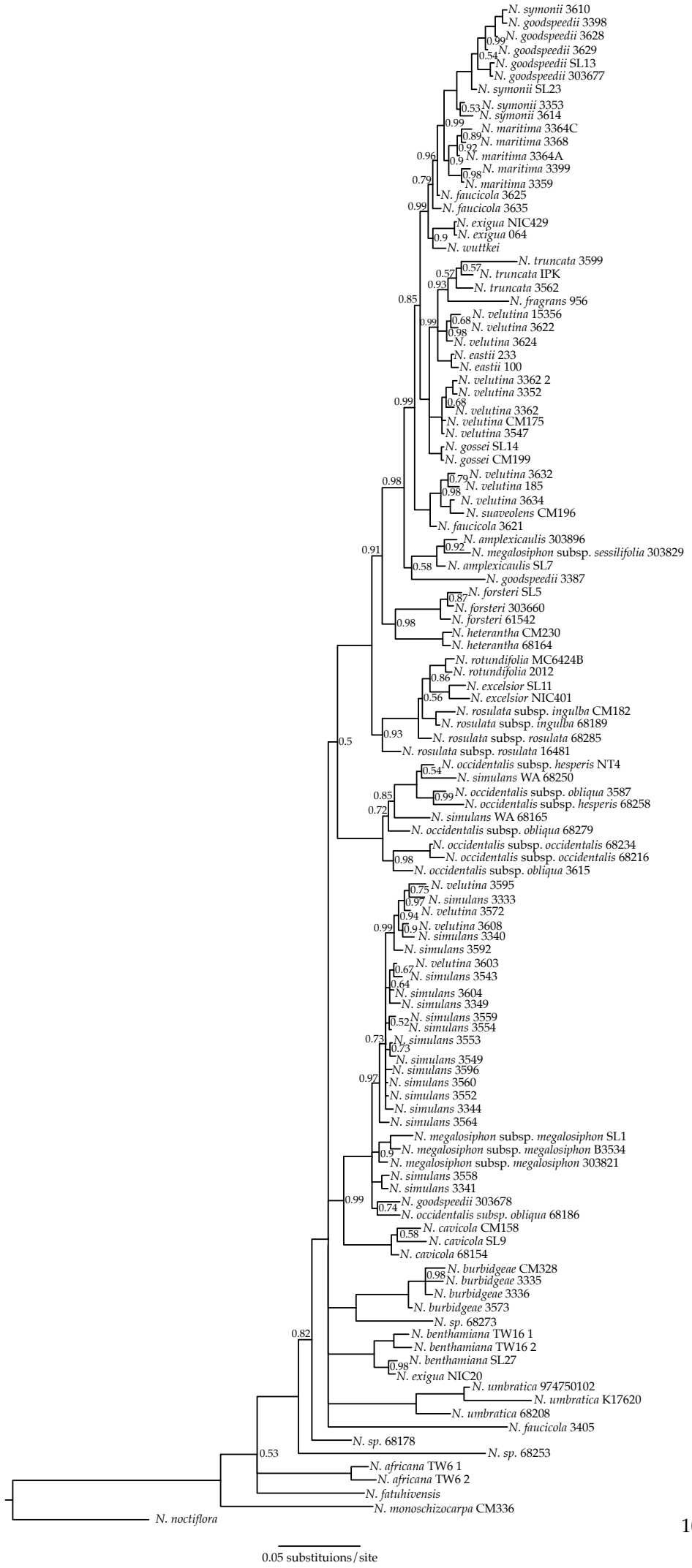
← **Figure 4.3** Plastome tree showing the posterior probability consensus from the PhyloBayes analysis. Nodes show posterior probabilities >0.5; nodes without values have posterior probabilities of 1.00. Clades with <0.5 PP are collapsed. Note *N. monoschizocarpa* and *N. noctiflora* (outgroup) branches have been truncated: inset shows the full branch lengths.

rDNA tree

The rDNA trees exhibit resolution of terminal clades, although the backbone of the Australian clade is not fully resolved (Figure 4.4). Polytomies also exist at a lower level, but this is due to the lack of variation in rDNA between populations. The first diverging lineage in section *Suaveolentes* is found to be *N. monoschizocarpa*, followed by *N. fatuhivensis* and *N. africana* that form an unresolved clade with low support (PP 0.62). There is high support (PP 0.99) for the rest of the Australian species as a clade. The first diverging taxa within this Australian clade are the potentially new taxa *N.* sp. 68178 and *N.* sp. 68253 from Western Australia. This is followed by a core clade of Australian taxa (PP 0.99) the backbone of which is largely unresolved. *Nicotiana faucicola* 3405 is found by itself in this clade, and several small species-specific clades are found with high support but for which the relationships to one another are unknown: *N. umbratica* (PP 1.00); *N. benthamiana* (PP 1.00); *N. burbidgeae* (PP 1.00); *N. occidentalis* s.l. (PP 1.00). A medium-sized clade is found to comprise all of the *N. simulans* populations, *N. megalosiphon* subsp. *megalosiphon* and a few difficult taxa, with *N. cavicola* being sister to the rest of this clade.

The final large clade has more resolution in the backbone of the tree. The first diverging subclade consists of *N. rosulata* subsp. *rosulata*, *N. rosulata* subsp. *ingulba*, *N. rotundifolia* and *N. excelsior*, of which the last two species are sister. The next clade consists of *N. forsteri* and *N. heterantha* as sister taxa. *Nicotiana megalosiphon* subsp. *sessilifolia* is found within a clade of *N. amplexicaulis*. *Nicotiana faucicola* 3621 is sister to a clade of *N. velutina* populations that also

includes *N. suaveolens*. The next medium-sized clade contains *N. gossei*, two subclades consisting of *N. velutina* populations (each with high support), the secondary polyploid taxon *N. eastii*, and *N. truncata* and *N. fragrans* as sister species (*N. truncata* populations forming a clade with relatively low support – PP 0.58). The final clade, sister to this medium-sized clade, puts *N. wuttkei* and *N. exigua* as sister species. *Nicotiana faucicola* 3635 and 3625 populations are then successively sister to the remainder of the clade. All *N. maritima* populations form a clade with reasonably high support (PP 0.89), which is then sister to a clade comprising *N. symonii* and *N. goodspeedii* populations. Three of four *N. symonii* populations form lineages that are then sister to all *N. goodspeedii* populations; the fourth *N. symonii* population is found nested within *N. goodspeedii* populations.


N. symonii 3610
N. goodspeedii 3398
N. goodspeedii 3628
N. goodspeedii 3629
N. goodspeedii SL13
N. goodspeedii 303677
N. symonii SL23
N. symonii 3353
N. symonii 3614
N. maritima 3364C
N. maritima 3368
N. maritima 3364A
N. maritima 3399
N. maritima 3359
N. faucicola 3625
N. faucicola 3635
N. exigua NIC429
N. exigua 064
N. wuttkei
N. truncata 3599
N. truncata IPK
N. truncata 3562
N. fragrans 956
N. velutina 15356
N. velutina 3622
N. velutina 3624
N. eastii 233
N. eastii 100
N. velutina 3362 2
N. velutina 3352
N. velutina 3362
N. velutina CM175
N. velutina 3547
N. gossei SL14
N. gossei CM199
N. velutina 3632
N. velutina 185
N. velutina 3634
N. suaveolens CM196
N. faucicola 3621
N. amplexicaulis 303896
N. megalosiphon subsp. sessilifolia 303829
N. amplexicaulis SL7
N. goodspeedii 3387
N. forsteri SL5
N. forsteri 303660
N. forsteri 61542
N. heterantha CM230
N. heterantha 68164
N. rotundifolia MC6424B
N. rotundifolia 2012
N. excelsior SL11
N. excelsior NIC401
N. rosulata subsp. ingulba CM182
N. rosulata subsp. ingulba 68189
N. rosulata subsp. rosulata 68285
N. rosulata subsp. rosulata 16481
N. occidentalis subsp. hesperis NT4
N. simulans WA 68250
N. occidentalis subsp. obliqua 3587
N. occidentalis subsp. hesperis 68258
N. simulans WA 68165
N. occidentalis subsp. obliqua 68279
N. occidentalis subsp. occidentalis 68234
N. occidentalis subsp. occidentalis 68216
N. occidentalis subsp. obliqua 3615
N. velutina 3595
N. simulans 3333
N. velutina 3572
N. velutina 3608
N. simulans 3340
N. simulans 3592
N. velutina 3603
N. simulans 3543
N. simulans 3604
N. simulans 3349
N. simulans 3559
N. simulans 3554
N. simulans 3553
N. simulans 3549
N. simulans 3596
N. simulans 3560
N. simulans 3552
N. simulans 3344
N. simulans 3564
N. megalosiphon subsp. megalosiphon SL1
N. megalosiphon subsp. megalosiphon B3534
N. megalosiphon subsp. megalosiphon 303821
N. simulans 3558
N. simulans 3341
N. goodspeedii 303678
N. occidentalis subsp. obliqua 68186
N. cavicola CM158
N. cavicola SL9
N. cavicola 68154
N. burbidgeae CM328
N. burbidgeae 3335
N. burbidgeae 3336
N. burbidgeae 3573
N. sp. 68273
N. benthamiana TW16 1
N. benthamiana TW16 2
N. benthamiana SL27
N. exigua NIC20
N. umbratica 974750102
N. umbratica K17620
N. umbratica 68208
N. faucicola 3405
N. sp. 68178
N. sp. 68253
N. africana TW6 1
N. africana TW6 2
N. fatuhivensis
N. monoschizocarpa CM336
N. noctiflora
0.05 substituions/site

← **Figure 4.4** rDNA tree showing the posterior probability consensus tree from the PhyloBayes analysis. Nodes show posterior probabilities >0.5; nodes without indicated values have posterior probabilities of 1.00. Clades with <0.5 PP are collapsed.

Genomic repeats

The tree reconstructed from genomic repeat abundances is shown in Figure 4.5. Overall this tree has little resolution, with most of the Australian taxa forming a large polytomy. There is variation in the abundance of repeats between taxa, but almost all genomic repeats are shared in roughly similar amounts. There is support for *N. africana* as an early branching lineage, distinct from the rest of section *Suaveolentes*, and this can be seen clearly in the raw cluster abundance data.

Additionally, there are a few distinct groups in the repeat tree, which represent a small number of clade-specific repeats (and their abundances) in these groups. All three *N. forsteri* accessions form a clade (BP 80); two *N. burbidgeae* accessions form a clade (BP 60). There is support for a clade consisting of *N. truncata* and *N. fragrans* accessions (BP 90), and within this a subclade consisting of two *N. umbratica* accessions (BP 100).

The re-analysis of a subset of taxa with greater read numbers (i.e. one order of magnitude more data; genome proportion of 0.5% analysed instead of the initial ~0.05%) resulted in a similar topology with no resolution amongst the Australian species of *Nicotiana* section *Suaveolentes* (the core clade) and as such is not presented here.

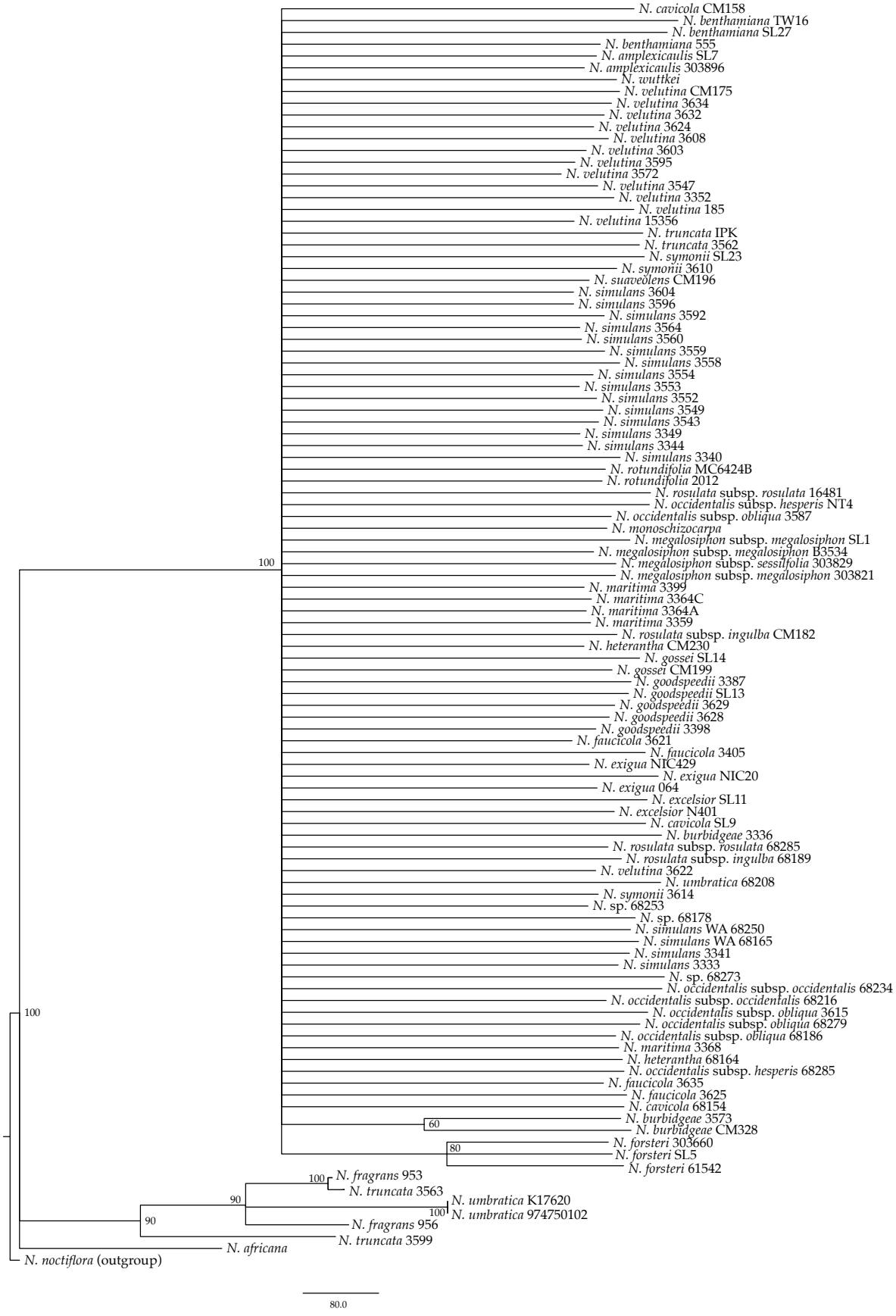


Figure 4.5 Phylogenetic tree based on repeat abundances. The top 1000 most abundant clusters were used with 100 symmetric bootstraps. Bipartition frequencies are printed on the single most parsimonious tree.

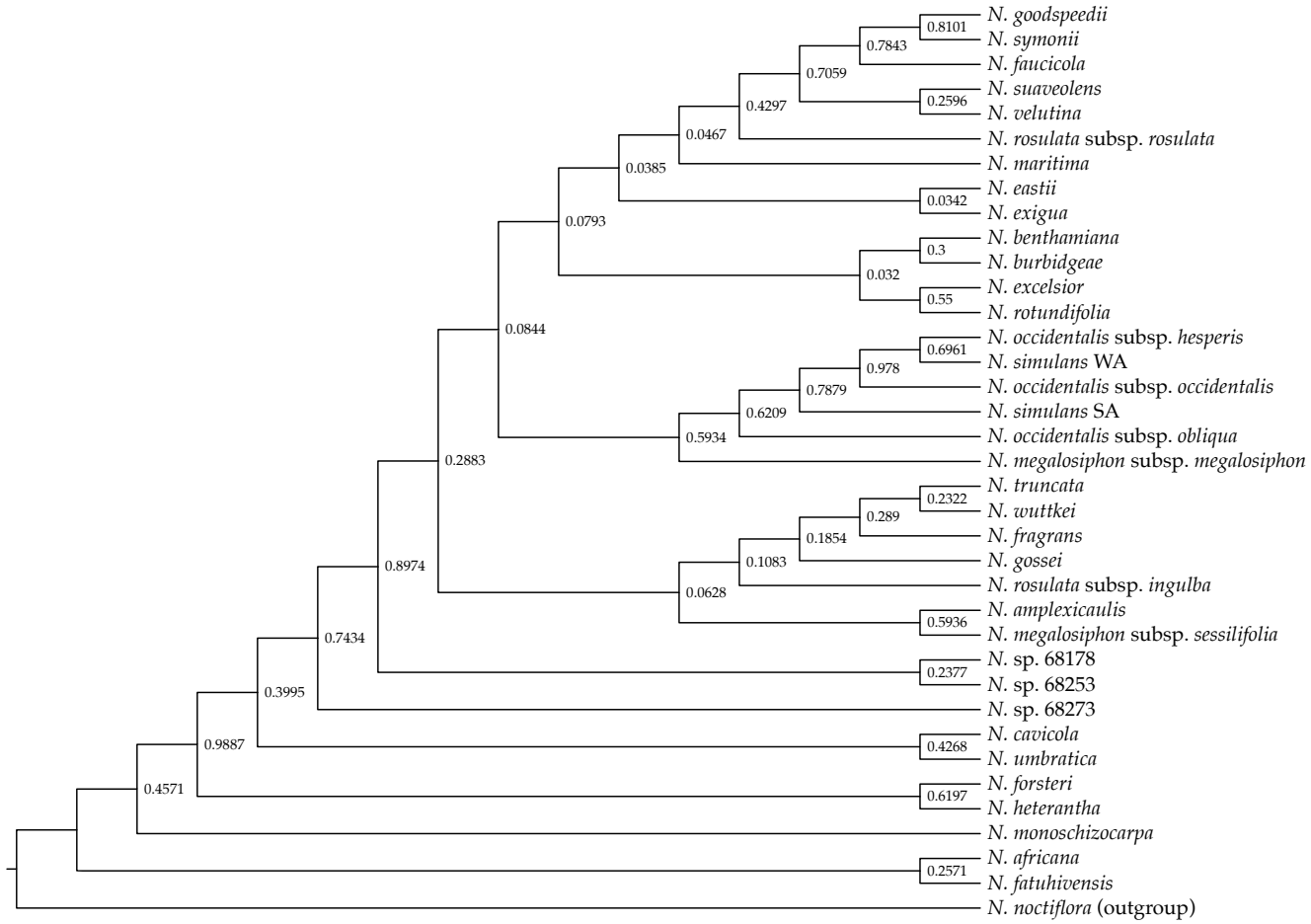
Species tree

The species tree analysis with *BEAST resulted in 19,820 trees post burn-in; the maximum clade credibility tree (maximising the product of clade posterior probabilities) is shown in Figure 4.6A. Overall the backbone of this tree is relatively poorly supported, thereby depicting the uncertainty in the species tree inference (see Figure 4.7) and the incongruence found between the plastome and rDNA data. However, there are some strongly support nodes, such as the Australian taxa minus *N. monoschizocarpa* (PP 0.99) and two nested clades of Australian taxa from *N.* sp. 68273 onwards (PP 0.74 and 0.90, respectively).

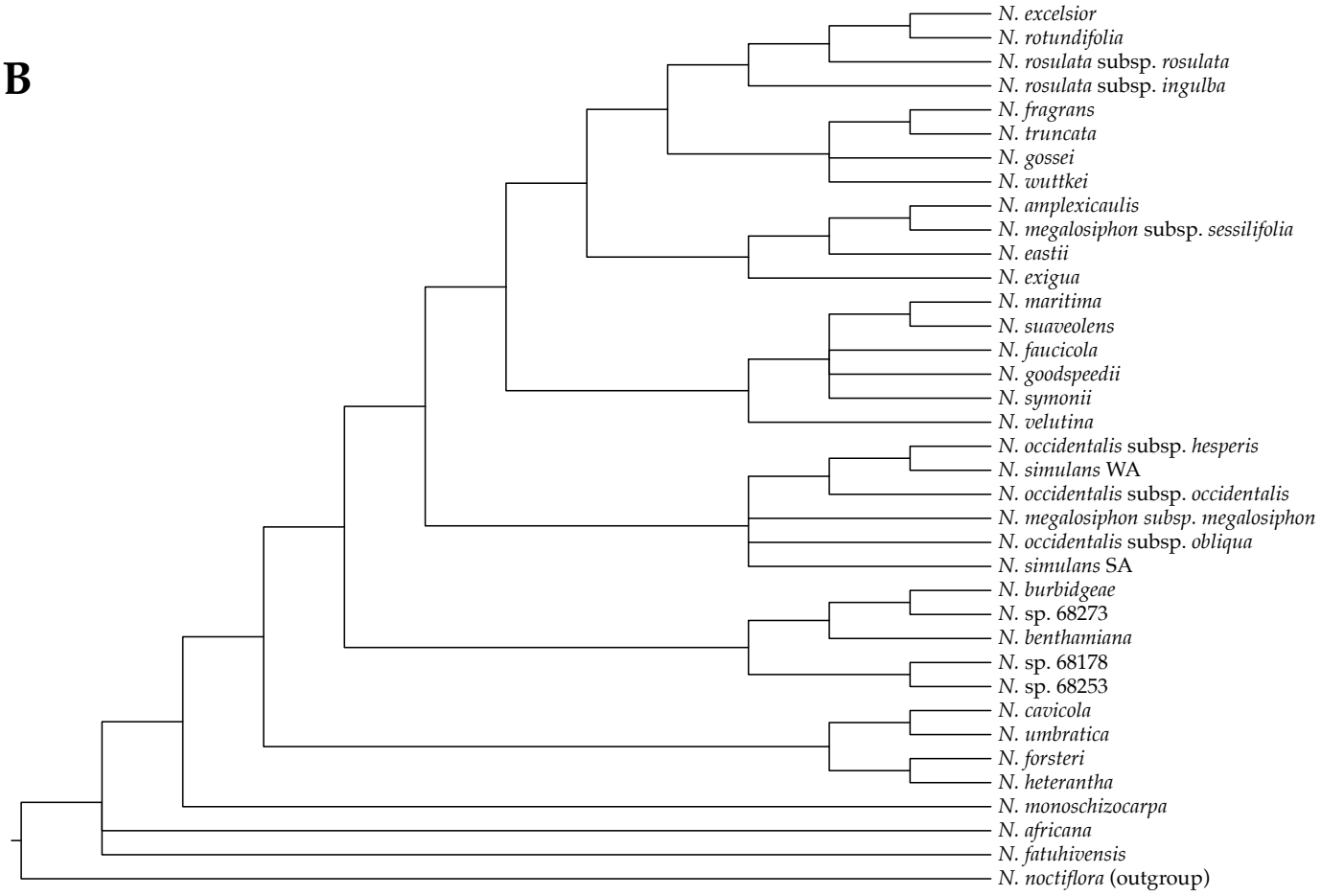
Despite this, there are some moderately supported subclades that consist of tightly-knit groups of taxa, for instance the clade comprising *N. goodspeedii*, *N. symonii*, *N. faucicola*, *N. suaveolens*, *N. velutina* (PP 0.71) and the clade comprising *N. occidentalis* (all subspecies), *N. simulans* (WA and SA forms) and *N. megalosiphon* subsp. *megalosiphon* (PP 0.59). There are also some moderately well-supported sister species relationships: *N. goodspeedii* and *N. symonii* (PP 0.81); *N. excelsior* and *N. rotundifolia* (PP 0.55); *N. simulans* SA and *N. occidentalis* (PP 0.79); *N. forsteri* and *N. heterantha* (PP 0.62).

The deep coalescence analysis in Mesquite identified 48 best trees with between 200-204 deep coalescences per tree. The strict consensus of these trees is presented in Figure 4.6B. The topology of this tree is very similar to the *BEAST tree, but with some differences particularly at lower levels, e.g. the sister species relationship of *N. fragrans* and *N. truncata* for which there is some evidence in the plastomes, rDNA and repeat analyses.

A


0.8101
0.7843
0.7059
0.2596
0.4297
0.0467
0.0385
0.0342
0.0793
0.3
0.032
0.55
0.0844
0.6961
0.978
0.7879
0.6209
0.5934
0.2883
0.2322
0.289
0.1854
0.1083
0.0628
0.8974
0.5936
0.7434
0.2377
0.3995
0.9887
0.4268
0.4571
0.6197
0.2571
N. goodspeedii
N. symonii
N. faucicola
N. suaveolens
N. velutina
N. rosulata subsp. rosulata
N. maritima
N. eastii
N. exigua
N. benthamiana
N. burbidgeae
N. excelsior
N. rotundifolia
N. occidentalis subsp. hesperis
N. simulans WA
N. occidentalis subsp. occidentalis
N. simulans SA
N. occidentalis subsp. obliqua
N. megalosiphon subsp. megalosiphon
N. truncata
N. wuttkei
N. fragrans
N. gossei
N. rosulata subsp. ingulba
N. amplexicaulis
N. megalosiphon subsp. sessilifolia
N. sp. 68178
N. sp. 68253
N. sp. 68273
N. cavicola
N. umbratica
N. forsteri
N. heterantha
N. monoschizocarpa
N. africana
N. fatuhivensis
N. noctiflora (outgroup)

B


N. excelsior
N. rotundifolia
N. rosulata subsp. rosulata
N. rosulata subsp. ingulba
N. fragrans
N. truncata
N. gossei
N. wuttkei
N. amplexicaulis
N. megalosiphon subsp. sessilifolia
N. eastii
N. exigua
N. maritima
N. suaveolens
N. faucicola
N. goodspeedii
N. symonii
N. velutina
N. occidentalis subsp. hesperis
N. simulans WA
N. occidentalis subsp. occidentalis
N. megalosiphon subsp. megalosiphon
N. occidentalis subsp. obliqua
N. simulans SA
N. burbidgeae
N. sp. 68273
N. benthamiana
N. sp. 68178
N. sp. 68253
N. cavicola
N. umbratica
N. forsteri
N. heterantha
N. monoschizocarpa
N. africana
N. fatuhivensis
N. noctiflora (outgroup)

← **Figure** 4.6 Species trees from (A) *BEAST multicoalescent analysis and (B) Mesquite minimising deep coalescences. Numbers on nodes in (A) represent posterior probabilities on the maximum clade credibility tree.

Figure 4.7 Species tree from *BEAST analysis visualised with DensiTree, showing all 19,820 trees overlaid and therefore the inherent uncertainty in the Bayesian estimate of the species tree.

Genealogical Sorting Index (GSI)

Results of the GSI analysis are presented in Table 4.3 for the plastome, rDNA and combined data. Only a couple of taxa were found unable to reject the null hypothesis of mixed ancestry in all or most of the three datasets – these were *N. occidentalis* subsp. *obliqua* and *N. rosulata* subsp. *rosulata*. This indicates that the populations/accessions of these taxa were not significantly different from random placement on the trees. In addition to this, despite rejection of mixed ancestry with various levels of significance (p <0.05-0.001) many other species had low GSI values overall. For example, *N. goodspeedii* and *N. faucicola* had particularly low GSI values across all three datasets. Out of the widespread taxa, *N. simulans* populations had a much greater level of exclusivity across the datasets (GSI of 0.72-1) compared to *N. velutina* (GSI of 0.38-0.59).

Table 4.3 Genealogical sorting index (GSI) values and significance values for rejection of the null hypothesis (mixed ancestry); n.s. indicates non-significant results. Note that species with only one representative could not be calculated and are therefore not shown.

Species	Plastome	rDNA	Both
N. africana	1**	1**	1**
N. amplexicaulis	1**	0.496*	0.748**
N. benthamiana	0.152 n.s.	0.661***	0.406***
N. burbidgeae	1***	1***	1***
N. cavicola	1***	1***	1***
N. eastii	1**	1**	1**
N. excelsior	1**	1**	1**
N. exigua	1***	0.210*	0.605***
N. faucicola	0.210*	0.179*	0.195**
N. forsteri	1***	1***	1***
N. goodspeedii	0.209*	0.209**	0.209**
N. gossei	1**	1**	1**
N. heterantha	1**	1**	1**
N. rosulata subsp. *ingulba*	0.160 n.s.	1**	0.580**
N. maritima	1***	1***	1***
N. megalosiphon subsp. *megalosiphon*	0.128 n.s.	1***	0.564***
N. occidentalis subsp. *obliqua*	0.075 n.s.	0.117 n.s.	0.096 n.s.
N. occidentalis subsp. *occidentalis*	0.701***	0.608***	0.655***
N. rosulata subsp. *rosulata*	0.093 n.s.	0.244*	0.168 n.s.
N. rotundifolia	1**	1**	1**
N. simulans SA	0.721***	1***	0.861***
N. simulans WA	0.069 n.s.	0.328*	0.198*
N. symonii	1***	0.413***	0.707***
N. truncata	0.128 n.s.	1***	0.564***
N. umbratica	1***	1***	1***
N. velutina	0.382***	0.590***	0.486***

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Ancestral state reconstructions

Life history strategies

The ancestral state reconstruction of life history strategy (annual vs. perennial) is shown in Figure 4.8. In both maximum parsimony and maximum likelihood analyses the root of section *Suaveolentes* is reconstructed as perennial. The early diverging lineages *N. africana* and *N. fatuhivensis* are long lived perennial species (indeed *N. africana* takes several years to reach flowering size, and *N. fatuhivensis* forms small tree-like shrubs). Little is known about the narrow endemic *N. monoschizocarpa* so this taxon was coded as unknown. Out of the likely early lineages within Australia, *N. forsteri* and *N. heterantha* are perennial, whereas *N. cavicola* and *N. umbratica* are annuals.

Most of the remaining Australian taxa are annuals (and indeed in several cases short-lived ephemerals). There are, however, several notable reversals to perennial life history in the Australian clade. The narrow endemic *N. burbidgeae* was found to have long tap roots and regrowth on the previous year's stems, thereby suggesting it is a long-lived perennial. The southern coastal *N. maritima* has populations with huge plants that have persisted for many years. The putative new species *N. faucicola* was found to have large plants with new growth on top of previous year's stems. Finally, *N. fragrans*, which is found in the South Pacific has a woody caudex and is therefore likely perennial.

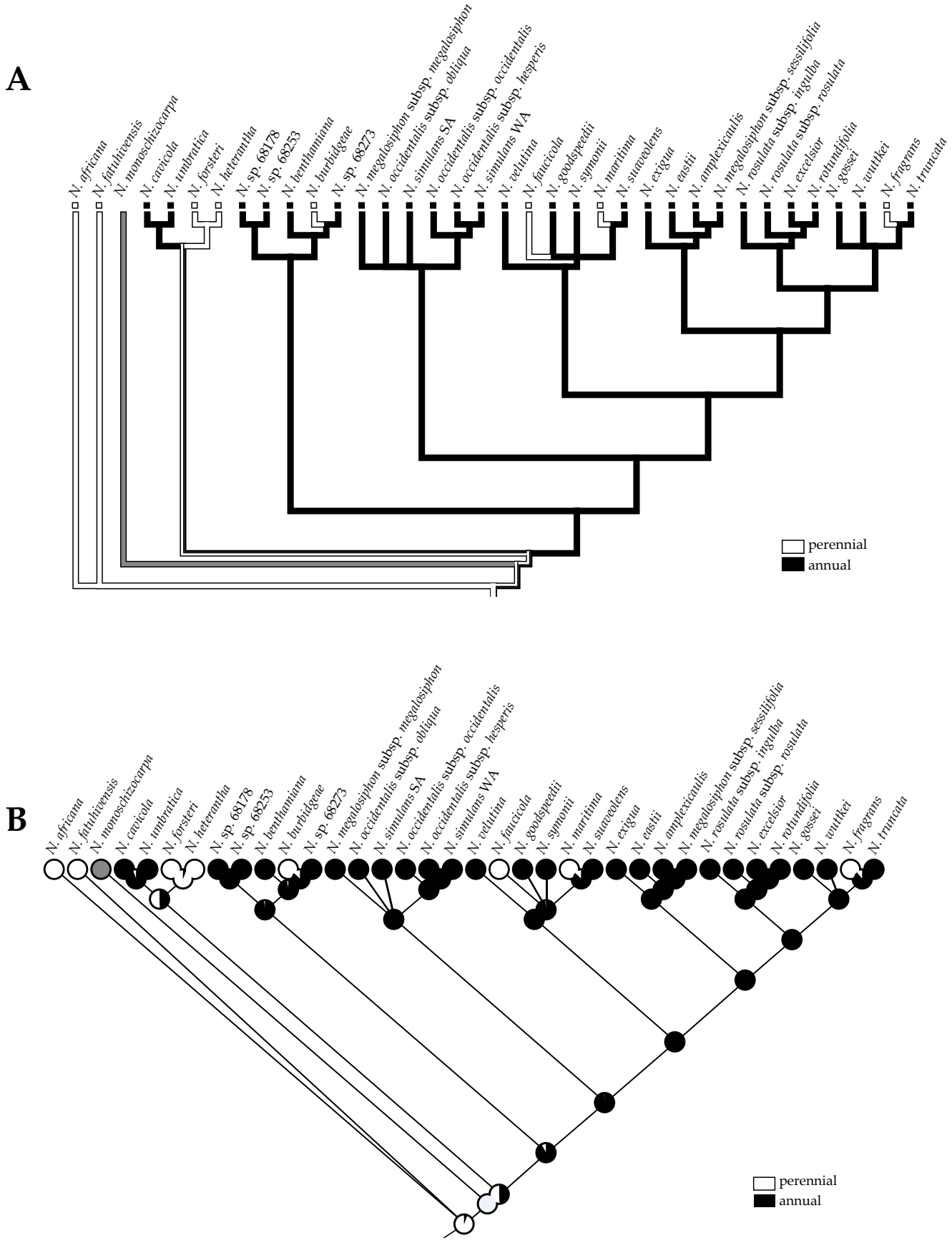
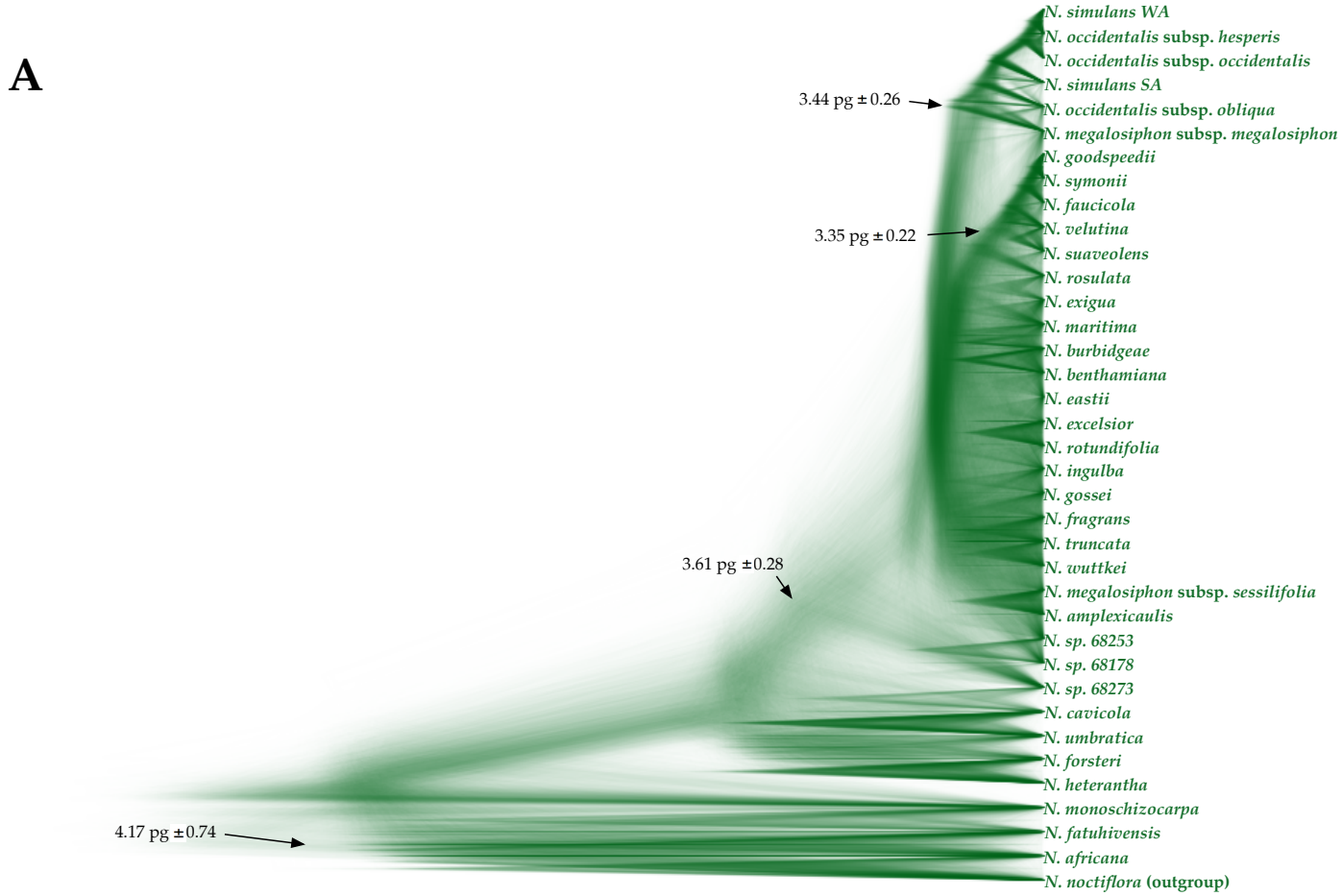
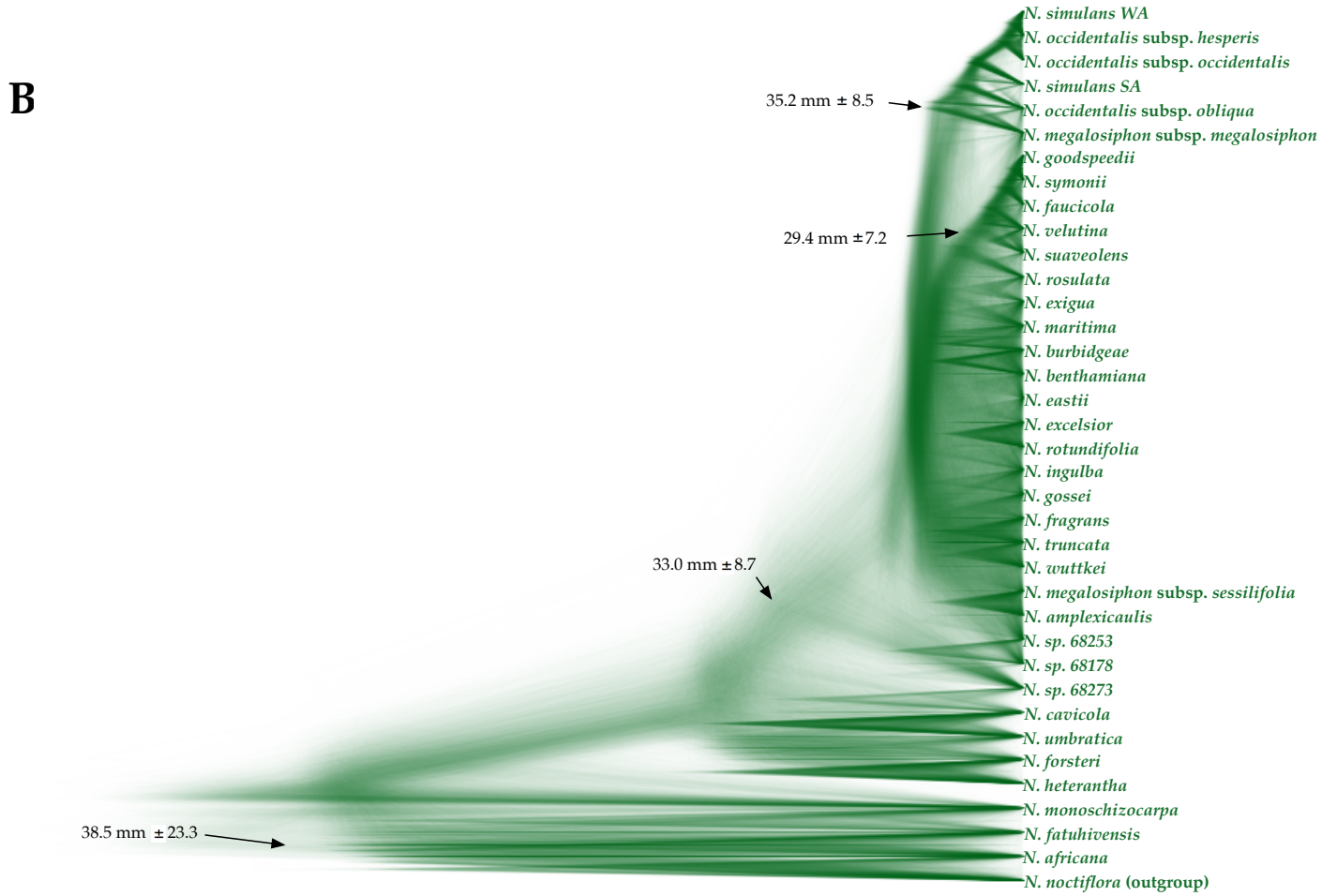


Figure 4.8 Ancestral reconstruction of life history strategy (annual vs. perennial) for *Nicotiana* section *Suaveolentes* using maximum parsimony (A) and maximum likelihood with the Mk1 model (B) in Mesquite.


A
3.44 pg ± 0.26
3.35 pg ± 0.22
3.61 pg ± 0.28
4.17 pg ± 0.74
N. simulans WA
N. occidentalis subsp. hesperis
N. occidentalis subsp. occidentalis
N. simulans SA
N. occidentalis subsp. obliqua
N. megalosiphon subsp. megalosiphon
N. goodspeedii
N. symonii
N. faucicola
N. velutina
N. suaveolens
N. rosulata
N. exigua
N. maritima
N. burbidgeae
N. benthamiana
N. eastii
N. excelsior
N. rotundifolia
N. ingulba
N. gossei
N. fragrans
N. truncata
N. wuttkei
N. megalosiphon subsp. sessilifolia
N. amplexicaulis
N. sp. 68253
N. sp. 68178
N. sp. 68273
N. cavicola
N. umbratica
N. forsteri
N. heterantha
N. monoschizocarpa
N. fatuhivensis
N. africana
N. noctiflora (outgroup)


B
35.2 mm ± 8.5
29.4 mm ± 7.2
33.0 mm ± 8.7
38.5 mm ± 23.3
N. simulans WA
N. occidentalis subsp. hesperis
N. occidentalis subsp. occidentalis
N. simulans SA
N. occidentalis subsp. obliqua
N. megalosiphon subsp. megalosiphon
N. goodspeedii
N. symonii
N. faucicola
N. velutina
N. suaveolens
N. rosulata
N. exigua
N. maritima
N. burbidgeae
N. benthamiana
N. eastii
N. excelsior
N. rotundifolia
N. ingulba
N. gossei
N. fragrans
N. truncata
N. wuttkei
N. megalosiphon subsp. sessilifolia
N. amplexicaulis
N. sp. 68253
N. sp. 68178
N. sp. 68273
N. cavicola
N. umbratica
N. forsteri
N. heterantha
N. monoschizocarpa
N. fatuhivensis
N. africana
N. noctiflora (outgroup)

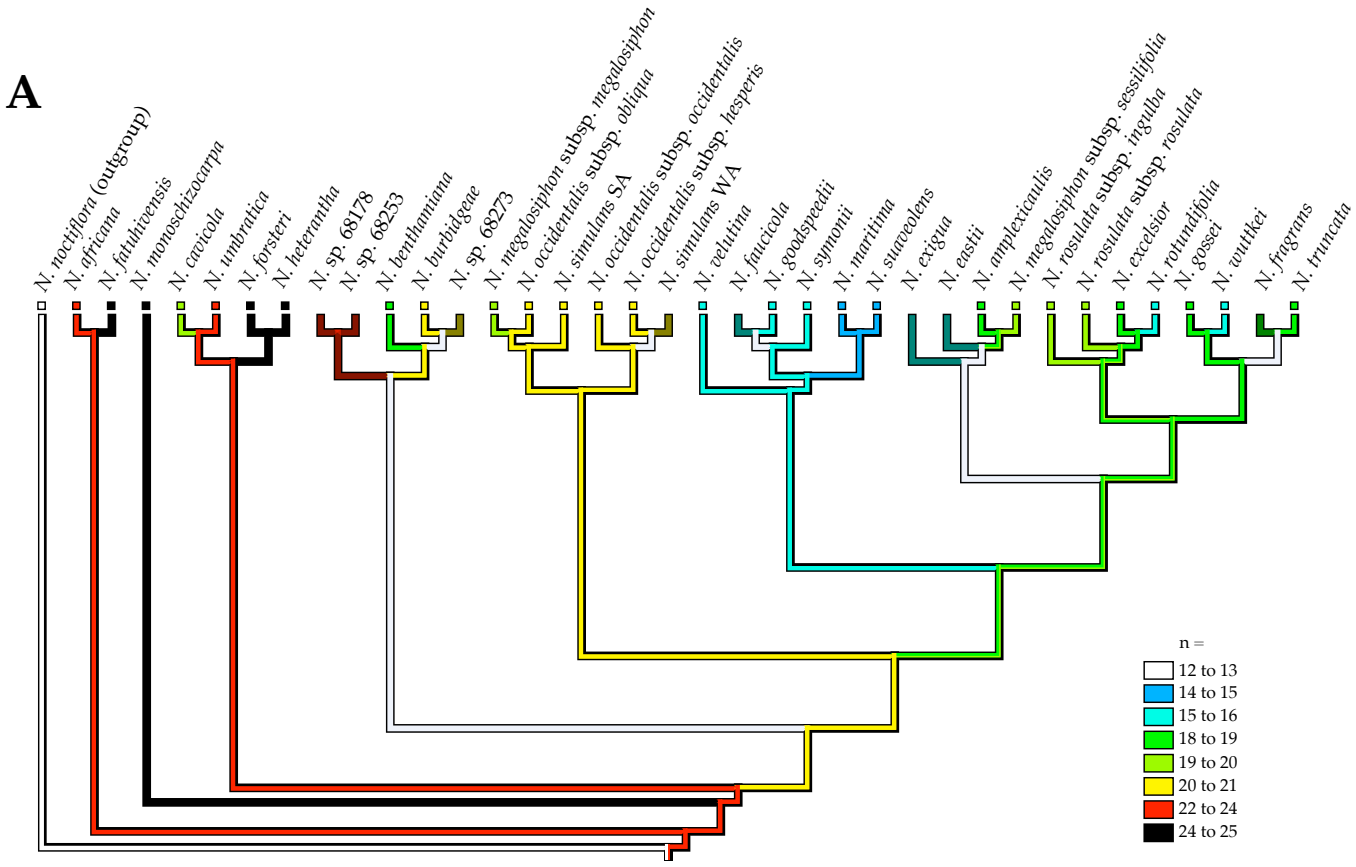
← **Figure 4.9** Ancestral state reconstruction of genome size (A) and corolla length (B) using the DensiTree of 19,820 species trees from the *BEAST analysis and a continuous random walk model of continuous character evolution in BayesTraits. Mean values at specific nodes are indicated with their standard deviation.

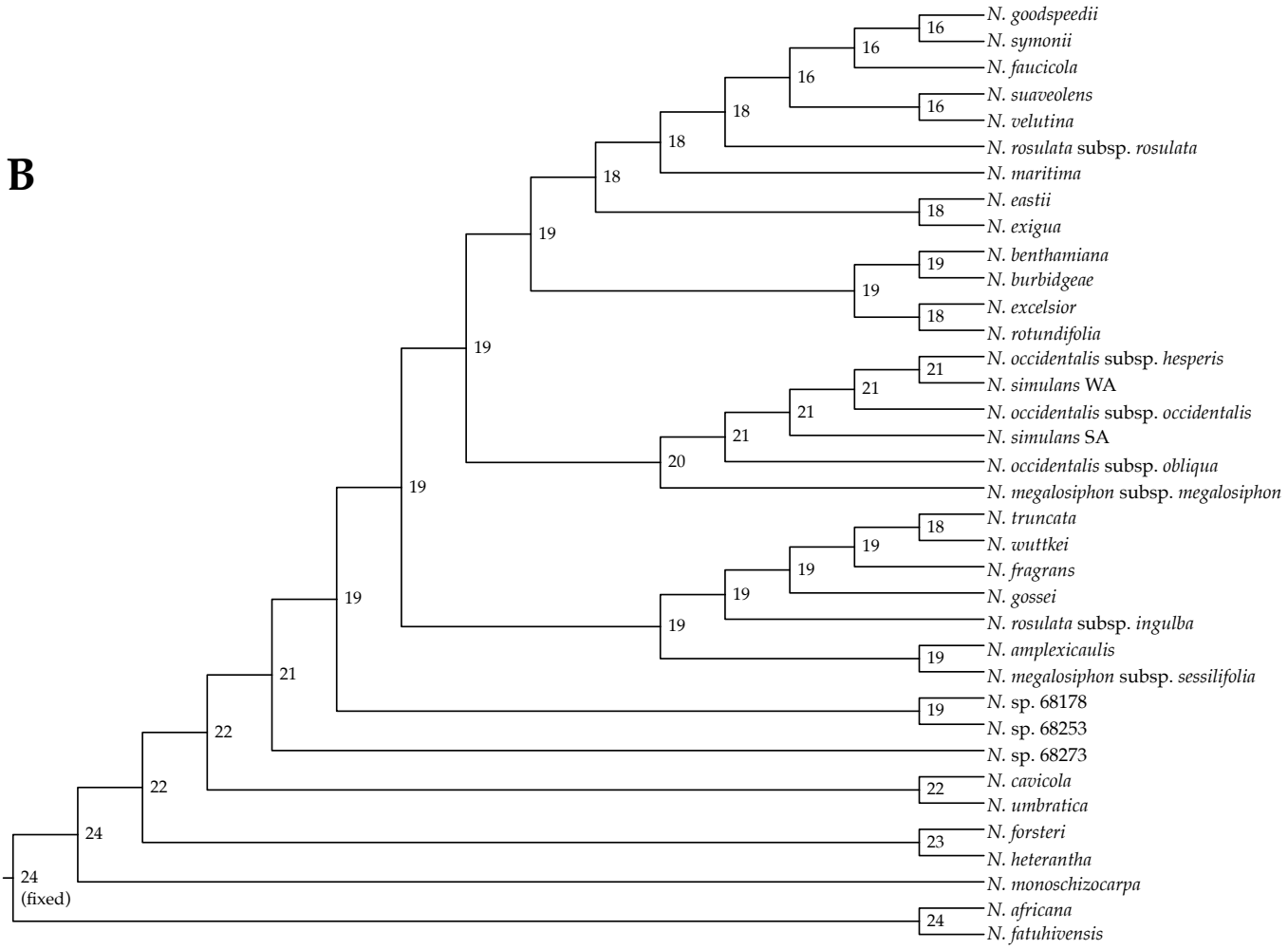
Genome size

The ancestral reconstruction of genome size suggests a value of ~4.17 pg for the root of *Suaveolentes* (Figure 4.9A). Further nodes are reconstructed roughly around 3.50 pg in most cases, which is in line with the average genome size for section *Suaveolentes* (3.66 pg). This indicates reasonable genome downsizing in the core Australia taxa of the section.

Corolla tube length

Corolla tube length is one of the most important morphological features for identifying the species of section *Suaveolentes*. The ancestral node was reconstructed at 38 mm (Figure 4.9B), and further nodes were reconstructed as between 29 and 35 mm.


A
N. noctiflora (outgroup)
N. africana
N. fatuhivensis
N. monoschizocarpa
N. cavicola
N. umbratica
N. forsteri
N. heterantha
N. sp. 68178
N. sp. 68253
N. benthamiana
N. burbidgeae
N. sp. 68273
N. megalosiphon subsp. megalosiphon
N. occidentalis subsp. obliqua
N. simulans SA
N. occidentalis subsp. occidentalis
N. occidentalis subsp. hesperis
N. simulans WA
N. velutina
N. faucicola
N. goodspeedii
N. symonii
N. maritima
N. suaveolens
N. exigua
N. eastii
N. amplexicaulis
N. megalosiphon subsp. sessilifolia
N. rosulata subsp. ingulba
N. rosulata subsp. rosulata
N. excelsior
N. rotundifolia
N. gossei
N. wuttkei
N. fragrans
N. truncata
n =
12 to 13
14 to 15
15 to 16
18 to 19
19 to 20
20 to 21
22 to 24
24 to 25


B
N. goodspeedii
N. symonii
N. faucicola
N. suaveolens
N. velutina
N. rosulata subsp. rosulata
N. maritima
N. eastii
N. exigua
N. benthamiana
N. burbidgeae
N. excelsior
N. rotundifolia
N. occidentalis subsp. hesperis
N. simulans WA
N. occidentalis subsp. occidentalis
N. simulans SA
N. occidentalis subsp. obliqua
N. megalosiphon subsp. megalosiphon
N. truncata
N. wuttkei
N. fragrans
N. gossei
N. rosulata subsp. ingulba
N. amplexicaulis
N. megalosiphon subsp. sessilifolia
N. sp. 68178
N. sp. 68253
N. sp. 68273
N. cavicola
N. umbratica
N. forsteri
N. heterantha
N. monoschizocarpa
N. africana
N. fatuhivensis
24 (fixed)

← **Figure 4.10** Ancestral reconstruction of chromosome number in *Nicotiana* section *Suaveolentes* using maximum parsimony in Mesquite (A) and (B) maximum likelihood in chromEvol using the *BEAST maximum clade credibility species tree.

Chromosome number

Chromosome number in section *Suaveolentes* is remarkably labile, ranging from n = 24 to n = 15 with almost every number in between. Ancestral reconstruction of chromosome number based on parsimony places the root of *Suaveolentes* as n = 24 (Figure 4.10A); the maximum likelihood analysis had the root fixed at n = 24. Descending dysploidy is a general pattern throughout the phylogenetic tree of section *Suaveolentes*, with a broad pattern of reduction from the root at n = 24 to n = 15 at the most derived parts of the tree. There is evidence for multiple independent drops in chromosome number, however, with at least two separate clades harbouring species with numbers n = 18 or lower (Figure 4.10).

Discussion

The origin of section Suaveolentes*: Diversification follows a lag phase post-polypoidisation*

The origin of *Nicotiana* section *Suaveolentes* approximately 6.8 Mya makes it the oldest of the polyploid sections in *Nicotiana*, confirming the relative estimates reported previously (e.g. Clarkson, 2007; Leitch *et al.,* 2008).

Section *Suaveolentes* diversifies into the majority of its current species diversity (~30 species) in mainland Australia within the last 2 mya, following a lag of approximately 3 my following the split of *N. forsteri* from the majority of Australian species (Figure 4.2). There is mixed support regarding the putative first lineage in Australia, as possibly *N. monoschizocarpa* or *N. forsteri*. *Nicotiana monoschizocarpa* was previously considered a subspecies of *N. forsteri* under its former name (*N. debneyi* subsp. *monoschizocarpa*). Although there is a morphological disparity between these taxa, there are similarities in habit and chromosome number (n = 24) and though *N. monoschizocarpa* is a narrow endemic from near Darwin (Northern Territory) it marginally overlaps with the range edge of *N. forsteri*.

Genome size evolution

The typical diploid genome size in *Nicotiana* (n = 12) is around 2.5 pg (Leitch *et al.,* 2008; http://data.kew.org/cvalues/). Although the exact parentage of section *Suaveolentes* is unclear due to the maternal progenitor being a putative homoploid hybrid (Kelly *et al.,* 2013), the typical size of a newly formed allopolyploid in *Nicotiana* is around 5 pg (e.g. *N. tabacum*, *N. rustica*). *Nicotiana africana* (n = 23) has a genome size of ~5.2 pg, which would be in line with the expected genome size of an initial polyploid lineage in *Nicotiana*. Out of the Australian species *N. forsteri* (n = 24) has close to this value at 4.8 pg. The rest of the section, and the Australian taxa, have undergone significant genome downsizing with most species having genome sizes of around 3.0-3.5 pg (Table

4.2). Genome downsizing is a common phenomenon post-polyploidisation and is associated with the removal of both genes and non-genic (i.e. repetitive) DNA. In *Nicotiana* relatively young allopolyploids have undergone less genome downsizing than older allopolyploids; e.g. *N. tabacum* (<0.2 My old) has a genome size of 5.18 pg compared to *N. nudicaulis* (~4.5 My old) with a genome size of 3.56 pg (Leitch *et al.*, 2008; Renny-Byfield *et al.*, 2011; Renny-Byfield *et al.*, 2013). In both these cases there has been removal of high-copy repetitive DNA, but in the case of *N. nudicaulis* this has been more extensive representing genome downsizing from expectation of approximately 19.2% (Renny-Byfield *et al.*, 2013). Section *Suaveolentes* is older than section *Repandae*, to which *N. nudicaulis* belongs, and as such it is expected that significant genome downsizing may have occurred. It is worthwhile noting that in early-branching clades downsizing is less extensive (or absent), and it appears that downsizing is associated with the later rapid diversification of *Suaveolentes* taxa within mainland Australia.

Plastome tree

The tree from plastome data is essentially fully resolved, with few nodes that have low support. The divergence of early lineages of the section in this analysis, i.e. the split of *N. africana*, *N. fatuhivensis* and also *N. forsteri*, are consistent with biological interpretations of the origin of the section and biogeographical aspects of their distributions. It has been hypothesised that the origin of the section in Australia is the result of long-distance dispersal (Clarkson *et al.*, 2004; Clarkson *et al.*, 2010; Marks, 2010; Kelly *et al.*, 2013). The ancestor of *Suaveolentes* could have evolved in South America (where the genus likely originated) and movement into Australia would likely have taken place across the southern Pacific. Thus the fact that the isolated *N. fatuhivensis* from the Marquesas Islands in the heart of the South Pacific Ocean forms an early lineage in the section provides support for this hypothesis. *Nicotiana forsteri* is distributed throughout eastern Australia, which would be the route into

Australia via the Pacific, and therefore support for *N. forsteri* as sister to the rest of the Australian species also provides evidence for this biogeographical origin of Australian *Suaveolentes* taxa. However, if *N. africana* is sister to the rest of section *Suaveolentes*, it is equally possible that the section dispersed across the Atlantic, reaching Australia subsequently. The occurrence of a significant lag between these initial lineages of section *Suaveolentes* and most of the speciation events is also clear in the plastome tree.

rDNA tree

The rDNA phylogenetic tree has less support in the backbone of the tree but more support for the monophyly of populations within certain species (e.g. *N. simulans*, *N. goodspeedii*, some groups of *N. velutina* populations).

Repeat tree

The analysis of genomic repeats resulted in a completely unresolved topology for section *Suaveolentes*. Despite some variation in genome size, it is clear that the differences in repeat abundance are not significant enough between species of Australian *Suaveolentes* to carry a phylogenetic signal. Although repeat abundances have been shown to group closely related cultivars of *Solanum lycopersicon* (Dodsworth *et al.*, 2015b) and resolve their relationship amongst other species of *Solanum*, the ability of this method to resolve recent, rapid radiations was until now unknown. As there is potential for repeat abundance differences to be useful at the intraspecific level, it may be that the rapid rate of evolution (as opposed to the low phylogenetic level) is what causes the failure of repeat abundances to have phylogenetic signal in *Nicotiana* section *Suaveolentes*. Thus incomplete lineage sorting would have resulted in ancestral polymorphism of repeats in the genomes of *Suaveolentes* taxa and as such these data are particularly deficient in phylogenetic signal.

Rampant non-monophyly of plastome haplotypes and incongruent datasets

At shallower depths there is a reasonable degree of congruence between the plastome and rDNA trees for section *Suaveolentes*. Despite this, the clear separation into mostly three main clades found in the plastome analysis was not found in the rDNA analysis. Instead the backbone of the tree is largely different, and several largely confusing results from the plastome tree made much more sense in light of the rDNA data (e.g. the close relationships of *N. goodspeedii* and *N. symonii* with near-exclusivity of population samples of each; the exclusivity of *N. truncata* populations that were otherwise widely dispersed in the plastome tree; the grouping of all *N. simulans* from South Australia (where the type originates) into a single clade, with no direct link of some accessions to the *N. occidentalis* clade).

In almost all cases, there is no evidence for hybrid taxa on morphological grounds, which would have potentially explained the rampant non-monophyly of populations in the plastome tree. Some accessions are almost certainly misidentified, and further study of these plants may reveal that species limits are not violated by the results obtained here. The lack of clustering of species based on the names provided could also be due to retention of ancestral polymorphism rather than hybridisation between species. This seems to involve most species within the three clades of Australian taxa (Figure 4.3), but clade I (*N. velutina*, *N. goodspeedii* etc.) seems to have a higher propensity for non-exclusivity. This might be because most of these species are reportedly n = 16, and thereby intercrossing of populations is expected to be more likely (and hybrids more viable) than between species differing in chromosome number. Thus plastid introgression could be a cause of this incongruence.

However, the species tree (Figure 4.7) also clearly displays uncertainty surrounding all of the Australian core clade, thereby invoking ancestral polymorphism and lineage sorting as a likely explanation for lack of resolution

in the data as a whole. Indeed, reciprocal non-monophyly of haplotypes is an expected stage in the diversification of lineages (Avise and Ball, 1990; Maddison, 1997; Schmidt-Lebuhn *et al.*, 2012). Over time genetic drift leads to lineage sorting and the extinction of haplotypes until haplotypes are reciprocally monophyletic. Particularly problematic species (in terms of lack of unique clustering) are often the most widespread ones (e.g. *N. velutina*, *N. goodspeedii*) and thus logically could represent a widespread gene pool from which more localised taxa have recently been derived. The GSI values (Table 4.3) are largely significant for these problematic taxa, which means that the null hypothesis of mixed ancestry could still be rejected despite overall low GSI values. This suggests that these species form independent lineages and does not provide a strong role for introgression. This would be consistent with insights from the field biology of these species, which suggests they are largely selfing and that populations are often isolated. Nonetheless hybridisation at homoploid and polyploid levels is frequent in *Nicotiana* and thus intrasectional introgression/hybridisation cannot be ruled out.

Origins and evolution of the Pacific taxa

Taxonomic clarification of the taxa that occur in the South Pacific was provided by Marks (2010a; 2010b). *Nicotiana debneyi* and *N. forsteri* are considered the same taxon, with the latter name having priority. *Nicotiana forsteri* is found on New Caledonia, Lord Howe Island, and in the eastern coast of Australia. The two remaining species found in the Pacific are *N. fragrans* and *N. fatuhivensis*, the latter elevated to species level (Marks, 2010b) from its placement as a variety of *N. fragrans* by Goodspeed (1954). *Nicotiana fragrans* occurs on New Caledonia and Tonga; *N. fatuhivensis* is endemic to the Marquesas Islands.

Morphologically these two species are similar in some respects – they both have relatively long corolla tubes (to over 70 mm), the presence of a woody caudex and perennial habit. Previous confusion (Marks, 2010a) seems to have taken place over the provenance of the accession used in molecular phylogenetic

studies of Clarkson *et al*. (2004; 2010); however, this was clearly *N. fatuhivensis* from the Marquesas (labelled variously as *N. fragrans*/*N. fragrans* var. *fatuhivensis*) and not *N. fragrans* from New Caledonia/Tonga.

The same accession of *N. fatuhivensis* (Wood 10529; Ua Huka – Marquesas Islands) was the only available material for this study, and the results presented here confirm its position as one of the earliest diverging taxa in section *Suaveolentes*. Three specimens of *N. fragrans* were sampled for gDNA, from BM herbarium specimens collected on Tonga and southern New Caledonia (Iles de Pines) – indeed two of these specimens were those studied morphologically by Marks (2010a; 2010b). One of these was successfully sequenced with Illumina technology, the other specimens were sequenced for nrITS and plastid markers to confirm the placement based on Illumina data (results not shown). Despite previous hypotheses regarding *N. fragrans* (Goodspeed, 1954; Clarkson *et al.*, 2004; Marks, 2010a; 2010b), *N. fragrans* is clearly nested within the Australian clade and therefore represents a secondary dispersal from mainland Australia. This is an unexpected result, yet given the morphological findings it is now clear that genetically *N. fragrans* and *N. fatuhivensis* represent two different evolutionary entities within section *Suaveolentes*. The chromosome count for *N. fragrans* of n = 24 (Marks, 2010a) has no provenance information, and thus is difficult to assign but is much more likely to belong to *N. fatuhivensis*.

Recurrent dysploidy in section Suaveolentes

The occurrence of dysploidy in section *Suaveolentes* has been long appreciated since initial studies of cytology and taxonomy in the genus that started almost a century ago, and it is one of the most enigmatic features of the section (Wheeler, 1935; Goodspeed, 1954). Nevertheless, the evolutionary context of dysploidy was hitherto unknown due to a lack of a reliable phylogenetic framework for the section. The results presented in this chapter show that chromosome number likely decreased multiple, independent times, from an ancestral ~n =

20-24, down to n = 18 in several species/clades (Figure 4.10). Goodspeed's hypothesis (1954) for these chromosome numbers in *Suaveolentes* was that two or three 'original' members, with numbers of n = 24 and n = 16, were involved in recurrent hybridisation in order to form the bulk of remaining species. Multiple hybridisations can be ruled out as a route of origin for the different chromosome numbers within the section. An important and complex role for intrasectional hybridisation seems unlikely given the fact that most species have a propensity for self-pollination. The most apparent feature of the phylogenetic hypotheses presented is the lack of variation due to recent and rapid evolution.

A second hypothesis was provided by Marks (2010a), suggesting that the phylogeny of the section in fact mirrors the drop in chromosome number, with ancestral species having much higher numbers than more derived ones. This hypothesis has some support from these analyses, but it is too simplistic. As shown by the phylogenetic analyses in this chapter, there are multiple independent instances of chromosome number reduction in *Suaveolentes*. Thus chromosome number reduction appears to have a recurrent role in the recent diversification of this group in different parts of Australia at roughly the same time. It may be that the process of chromosome reduction is related to local adaptation and formation of favourable linkage groups that are adaptive in the extreme environments in which these species grow.

Evolution of life history strategy in section Suaveolentes

In tandem with genome downsizing and chromosome number reduction there is a shift from a perennial habit to an annual one in section *Suaveolentes* (Figure 4.8). This likely reflects the adaptation of new taxa to the arid Eremaean zone of Australia, where the greatest species diversity is found today. Rapid cycling taxa include for example *N. simulans* and *N. truncata*, the former of which is found across the stony plains of South Australia and the latter in specific habitats in the same general area. What role these genomic factors have in the

ecological adaptation and rapid speciation of the group is currently unknown and a topic that requires much further investigation.

Overall these results taken together present a picture of *Nicotiana* section *Suaveolentes* as an ongoing recent and rapid radiation that has occurred post-polyploidisation.

Chapter 5 General Discussion

Publication information

Some of the ideas presented in this chapter are published in the following articles, for which I was the lead author. All co-authors contributed to discussions, edited and approved the final manuscripts.

Dodsworth S, Chase MW, Leitch AR. (2015) Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society*, DOI: 10.1111/boj.12357.

Dodsworth S, Leitch AR, Leitch IJ. (2015) Genome size diversity in angiosperms and its influence on gene space. *Current Opinion in Genetics and Development*, DOI: 10.1016/j.gde.2015.10.006.

Diversification follows a lag phase in Nicotiana *section* Suaveolentes

Nicotiana section *Suaveolentes* is a further example of the WGD radiation lag-time model, originally proposed by Schranz *et al*. (2012) and with recent statistical support garnered by Tank *et al*. (2015). This postulates that after a WGD event takes place, lineages often undergo increases in diversification rates, but only after a significant period of time (i.e. a lag), of several millions of years. In *Nicotiana* section *Suaveolentes* this lag phase appears to be in the order of ~4 million years, during which there were only a few speciation events after the arrival of this lineage in Australia. The section then rapidly diversifies within the last 2 million years, forming up to 30 species including new taxa, with clades independently undergoing dysploid chromosome reduction and adaptation to a wide variety of environmental conditions.

In tandem with the speciation events themselves clearly involving morphological, ecological and genomic changes, all species in the core Australian clade (i.e. recently diversified species) have undergone significant genome downsizing, from a hypothetical ancestral 4.2 pg to between 2.8 and 3.6 pg. This is a complex result of diploidization processes that have simultaneously removed repetitive DNA and genes, whilst restructuring the chromosomes from an ancestral n = 24 down to n = 15 in some species. Further research is needed to tackle which aspects of this diploidization process are key to the diversification and adaptation of the different clades within sect. *Suaveolentes*.

The lag phase post-polyploidisation and pre-diversification observed in *Nicotiana* section *Suaveolentes* is a phenomenon that has been found across angiosperms at various phylogenetic levels, resulting in a series of nested radiations (Tank *et al*., 2015). These findings reshape the recent debate on the importance of polyploidy in plants (see for example, Gorelick and Olson, 2013;

Mayrose *et al*., 2011; Soltis *et al*., 2014; Wood *et al*., 2009). Given that the immediate effects of polyploidisation can often be detrimental due to 'genomic shock' (McClintock, 1984), along with reduced fertility and a bloated genome that is inefficient to selection with a potential ecological cost (Chester *et al*., 2012; Conant *et al*., 2014; Husband, 2000; Leitch and Leitch, 2008; Levin, 1975; Guignard *et al*., submitted; Neiman *et al*., 2013; Otto, 2007; Stebbins, 1971; Šmarda *et al*., 2013; Yant *et al*., 2012), it begs the question as to why angiosperms have experienced so many rounds of polyploidy in their ancestry (Bowers *et al*., 2003; Blanc and Wolfe, 2004; Van de Peer *et al*., 2009; Jiao *et al*., 2011). Polyploids are clearly not 'evolutionary dead-ends' (Stebbins, 1950) in the long term. It is likely that the benefits of polyploidy, i.e. genetic and genomic variation (Soltis and Soltis, 2000; Leitch and Leitch, 2008; Flagel and Wendel, 2009), can only be harnessed after some of the negative effects have been eliminated through the process of diploidisation (Dodsworth *et al*., 2015d). Australian soils are particularly low in phosphorus, which will be limiting (Vitousek *et al*., 2010) – however whether this can impose selection on a smaller genome at this scale of genome size is something that requires further study (Guignard *et al*., submitted; Greilhuber and Leitch, 2013).

One of the unusual phylogenetic results was the finding that *N. fragrans* is nested within the Australian taxa and therefore represents a secondary dispersal to the South Pacific much later than the origin of the section. What is also particularly surprising is the potential for a sister species relationship between *N. fragrans* and *N. truncata*, the latter being a narrow endemic found only near Oodnadatta in South Australia. This potential relationship is supported by plastomes, rDNA and genomic repeats. The ecological niche models (using 19 Bioclim variables) and photographs of habit are provided in Figure 5.1 for comparison. There may be some similarity in general habit between these species with basal rosettes and fairly fleshy leaves.

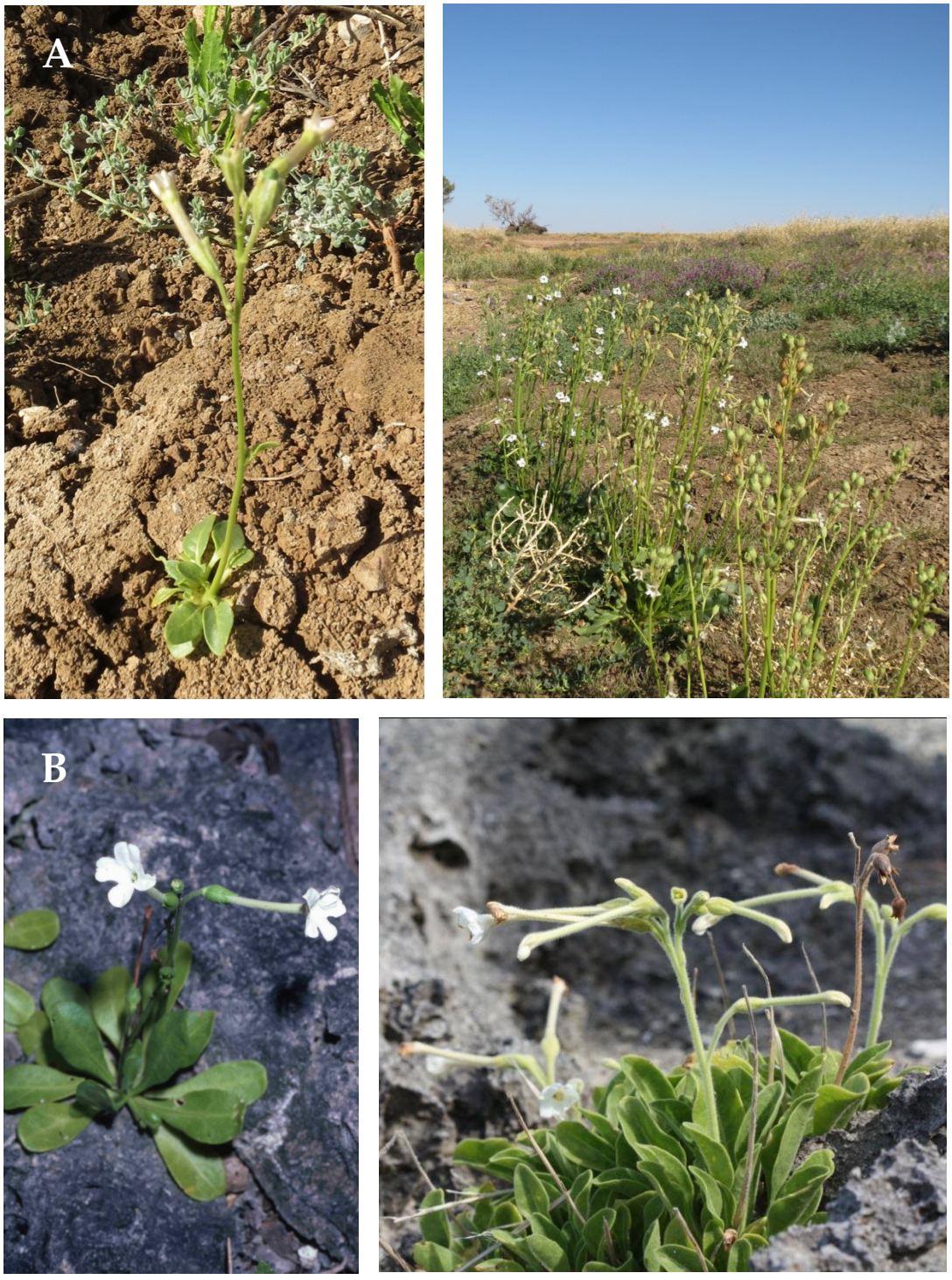


Figure 5.1 Comparison of *N. fragrans* and *N. truncata*. Photos of the general habit of each species. Row A – *N. truncata*; B – *N. fragrans*.

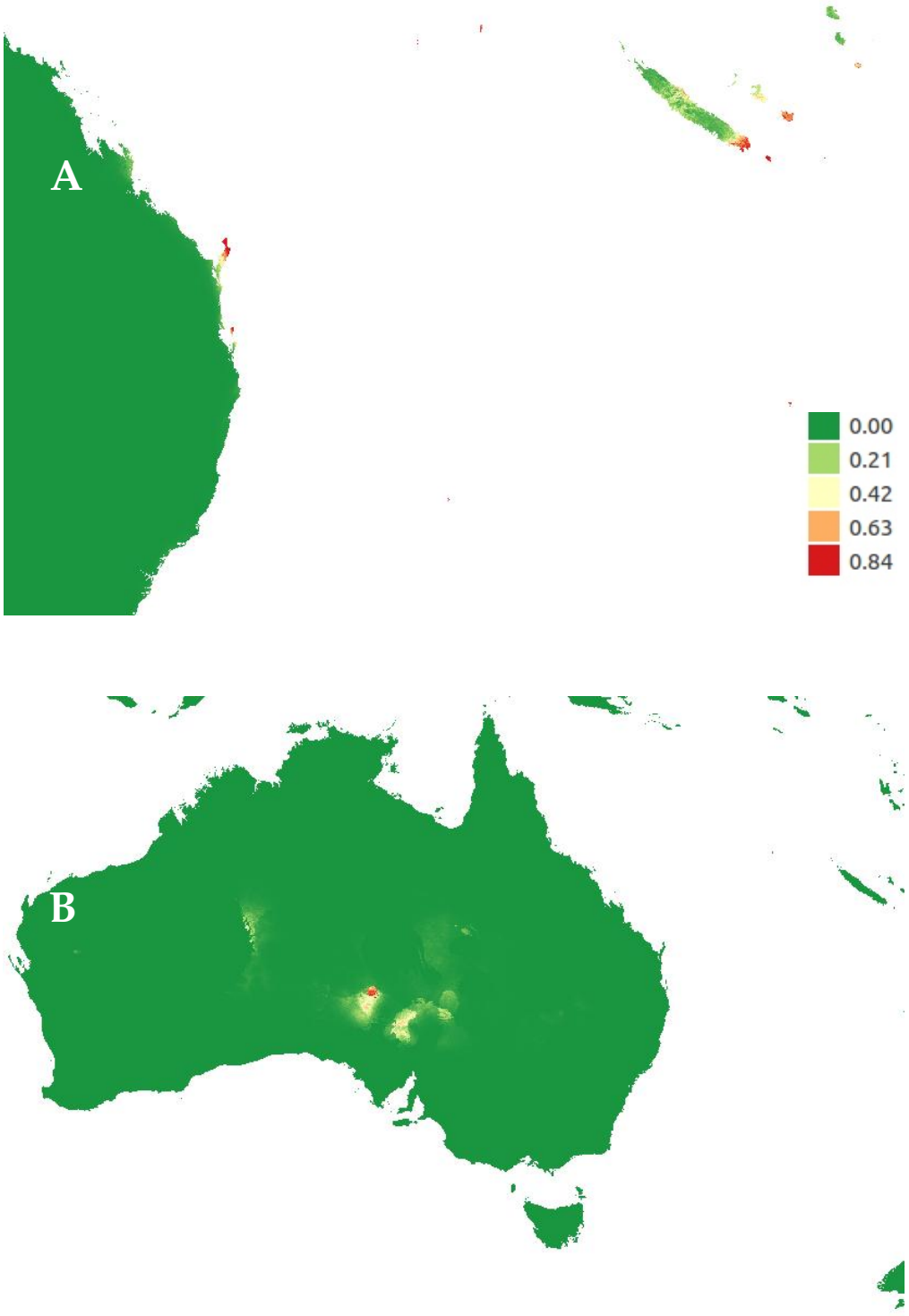


Figure 5.1 *cont.* Comparison of *N. fragrans* (A) and *N. truncata* (B). Ecological niche models based on distribution data from the Australian Virtual Herbarium and 19 Bioclim variables in MaxEnt. Note the narrow and disjunct distributions of these sister species. Coloured according to probability of occurrence (legend).

Taxonomy of section Suaveolentes—*new and cryptic species*

A potential new species initially collected by David Symon was identified in Claire Marks' thesis (2010) as sp. nov. 'Corunna' from Corunna Station near Iron Knob in South Australia; in this thesis it is denoted as *N. symonii* ined. The distinctive morphology of this species was confirmed by Marks (2010) in her morphological analyses of section *Suaveolentes*, suggesting that this taxon was worthy of recognition as a new species. Several collections were made of this species during field trips in 2013 and 2014 from the (potential) type locality and surrounding areas of the Corunna Hills. In the plastome tree all accessions of *N. symonii* form a highly supported clade, entirely separate from *N. goodspeedii* to which it has morphological similarities. In the rDNA tree, there is a close relationship between *N. symonii* and *N. goodspeedii*, with at least one population of *N. symonii* nested within *N. goodspeedii* populations. These data taken as a whole suggest the potential of a close relationship between these two taxa but that *N. symonii* does indeed represent a distinct lineage, perhaps derived by hybridisation between *N. goodspeedii* and *N. rosulata*.

Several putatively new taxa were found during recent fieldwork in August 2015 in northern Western Australia (M. Chase, M. Christenhusz, personal communication). These are included in the trees as *N.* sp. 68178, *N.* sp. 68253 and *N.* sp. 68273. In plastome, rDNA and species tree analyses they were found in interesting positions distant from the species they would otherwise key out as, e.g. *N. umbratica*, *N. rosulata* subsp. *rosulata* and *N. rotundifolia*, respectively. These taxa clearly represent novelties in the section, and this could indeed be indicative of further new taxa that are yet to be discovered, also signifying the need for much more extensive population-level analyses to delimit genetic entities that can then be recognised as species.

Taxonomy of section Suaveolentes—*intraspecific taxa*

The phylogenetic results presented in this thesis have several implications for intraspecific taxa within the section, some of which probably need to be elevated (back) to species level and others that can be considered as less useful intraspecific divisions of otherwise reasonable species. Whilst broadly it may appear that the three subspecies of *N. occidentalis* (subsp. *occidentalis*, *hesperis*, *obliqua*) were found to comprise a clade, this *N. occidentalis* complex requires much further work. The western Australian *N. simulans* is clearly nested within the *N. occidentalis* complex, which makes sense morphologically as it is very sticky and viscid (along with *N. occidentalis* subsp. *occidentalis* in particular). *N. occidentalis* subsp. *obliqua* from SA is potentially a form of *N. occidentalis* subsp. *occidentalis*, whereas the more widespread *N. occidentalis* subsp. *obliqua* from WA represents a distinct lineage. One *N. simulans* from WA is nested within this complex and could represent a misidentification of *N. occidentalis* subsp. *occidentalis*; the more typical collection of *N. simulans* from WA (68165) is clearly distantly related. Further investigations are needed at the population level for these WA taxa in order to define whether some of these are taxa that should be elevated to species level, and in order to define the limits of the *N. occidentalis* subspecific taxa.

The two subspecies of *N. rosulata* (subsp. *rosulata* and *ingulba*) are distinct morphologically, and *N. ingulba* was originally described as a species later sunk as a subspecies of *N. rosulata*. In the plastome tree these two subspecies are distinct, and in both cases (for *N. rosulata* subsp. *rosulata* and *N.* subsp. *ingulba*) accessions occur separated in the tree. Despite this, in the rDNA tree both subspecies are found close together in one clade, with the two accessions of *N. rosulata* subsp. *ingulba* forming a highly supported clade, and the two accessions of *N. rosulata* successively sister to the clade comprising *N. rosulata* subsp. *ingulba* amongst other taxa. Given this body of evidence it seems best to recognise *N. rosulata* and *N. ingulba* at the species level.

The two subspecies of *N. megalosiphon* are a case that requires further investigation. *Nicotiana megalosiphon* subsp. *megalosiphon* accessions form a highly supported clade in the rDNA tree, but in the plastome tree only two of the accessions are found together and one is found in a disparate clade. In both plastome and rDNA analyses the single accession of *N. megalosiphon* subsp. *sessilifolia* is found separate from the *N. megalosiphon* subsp. *megalosiphon* accessions. This indicates that *N. megalosiphon* subsp. *sessilifolia* is a different taxon, best given species-level recognition. Indeed, the distributions of these two subspecies do not appreciably overlap. This will require further sampling at the population level, together with further character sampling at the genomic level in order to definitively propose these subspecies as separate evolutionary lineages.

Genome size and genomic repeat evolution in section Suaveolentes

The phylogenetic tree based on genomic repeat abundances found no resolution amongst the species of *Suaveolentes*. Rather than a failing of the method, this astonishing lack of clear phylogenetic signal hints at biological phenomena underlying the radiation of the section. Despite this species are found on long branches of differing lengths, indicating some variation in repeat abundances; it may be that further investigation of different types of repeats (*sensu* Dodsworth *et al.*, 2015a) may yield different (and potentially more useful) signals. Most species have genome sizes of around ~3.5 pg, have undergone genome downsizing and experienced chromosome reorganisation via dysploid reduction to produce an array of lower chromosome numbers. The majority of species in section *Suaveolentes* likely diversified from a common ancestor with a genome size of ~3.5 pg with a lower chromosome number, as supported by the ancestral state reconstructions. Speciation then likely occurred rapidly, which led to a bewildering array of species that despite being morphologically and ecologically distinct entities, exhibit a similar genome size, broadly similar

repetitive element landscape of the genome and similarly low chromosome numbers. As such the abundance (and type) of different genomic repeats is similar across the species of *Suaveolentes*.

This apparent static nature of genome size and repeat profiles is particularly significant in the context of the chromosome number in *Suaveolentes*, as this could represent the formation of novel linkage groups that are involved in local adaptation and hence the speciation process (De Smet *et al*., 2013; Yeaman, 2013). One would expect that in many places the surrounding context of genes under selection (i.e. the repeats) would also have changed in these species, and these contextual changes could lead to expression differences in genes involved in adaptation, for example by *cis* modification to promoter regions of downstream genes (Dodsworth *et al*., 2015c; Grandbastien, 2015). Although broad-scale repeat dynamics appear relatively stationary in the species of section *Suaveolentes* (indicated by comparable size clusters representing different repeat families), it could be that finer-scale element accumulation/degeneration/deletion has an impact on gene evolution. The morphological and ecological differences of section *Suaveolentes* taxa could represent differences in expression, which would further add a role for fine-scale repetitive element changes to influence gene space (Dodsworth *et al*., 2015c).

In order to explore this fully a well-assembled genome sequence is required, which can then be used to map the context of particular repetitive elements and genes under selection in order to explore further these processes of adaptation in section *Suaveolentes*. Given the two draft genome sequences already available for *N. benthamiana* (Bombarely *et al*., 2012; Naim *et al*., 2012), the genome sequence availability of other *Nicotiana* species including *N. sylvestris* (the closest extant relative of one of the diploid progenitors of *Suaveolentes*), *N.*

tomentosiformis and *N. tabacum* (Sierro *et al*., 2013; 2014), it is likely that a better assembly of a *Suaveolentes* genome will be available in the near future.

A link between dysploidy, diploidisation and diversification

Genome downsizing is a common phenomenon post-polyploidisation in angiosperms (Leitch and Bennett, 2004; Meudt *et al*., 2015), largely as a result of a loss of genomic repeats (Leitch and Bennett, 2004; Lim *et al*., 2007; Renny-Byfield *et al*., 2011; Renny-Byfield and Wendel, 2014) and can therefore be considered as part of the diploidisation process. Genome sizes in angiosperms are skewed towards low values (Figure 5.2) and similarly so are chromosome numbers (Figure 5.3); neutral theories of the skew in genome sizes are inadequate (Oliver *et al*., 2007). Thus despite the propensity for polyploidisation in angiosperms, the overall pattern is for genomes to return to a diploid-like state in terms of overall genome size, chromosome number (Dodsworth *et al*., 2015d), and structural gene content (based on a huge body of literature on isozymes of 'diploid' species). Extensive chromosomal rearrangements are part and parcel of this process (e.g. Franzke *et al*., 2011). This is in stark contrast with the ferns, where such processes of diploidisation seem to not occur (Leitch and Leitch, 2012), and ferns include the highest chromosome count ever reported of 2n = c. 1440 in *Ophioglossum* (Abraham and Ninan, 1954). Linking this with recent work concerning diversification rates in angiosperms (Tank *et al*., 2015), it becomes apparent that the lag phase post-polyploidisation and prior to diversification is one associated with diploidisation and reorganisation of polyploid genomes. Polyploid genomes, as perhaps best exemplified by crop species, undergo a perplexing variety of reorganisation processes including fractionation, chromosome rearrangement and DNA loss (Wendel, 2015).

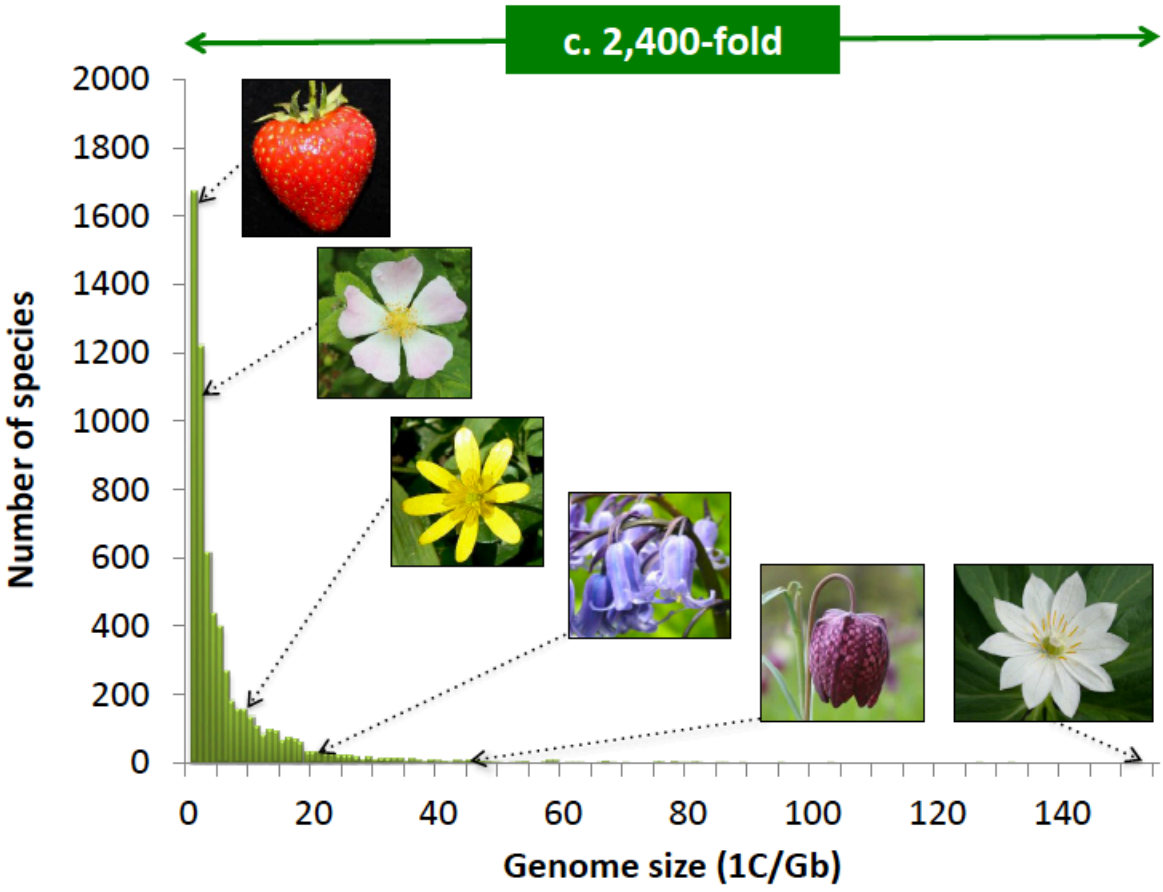


Figure 5.2 Genome size distribution in angiosperms based on 7542 C-values from the Plant DNA C-values database (www.data.kew.org/cvalues), adapted from (Dodsworth *et al.*, 2015c).

Nicotiana section *Suaveolentes* therefore represents an excellent case study for the link between diploidisation post-polyploidisation and lineage diversification in angiosperms, with both genome downsizing and chromosome number reduction readily apparent.

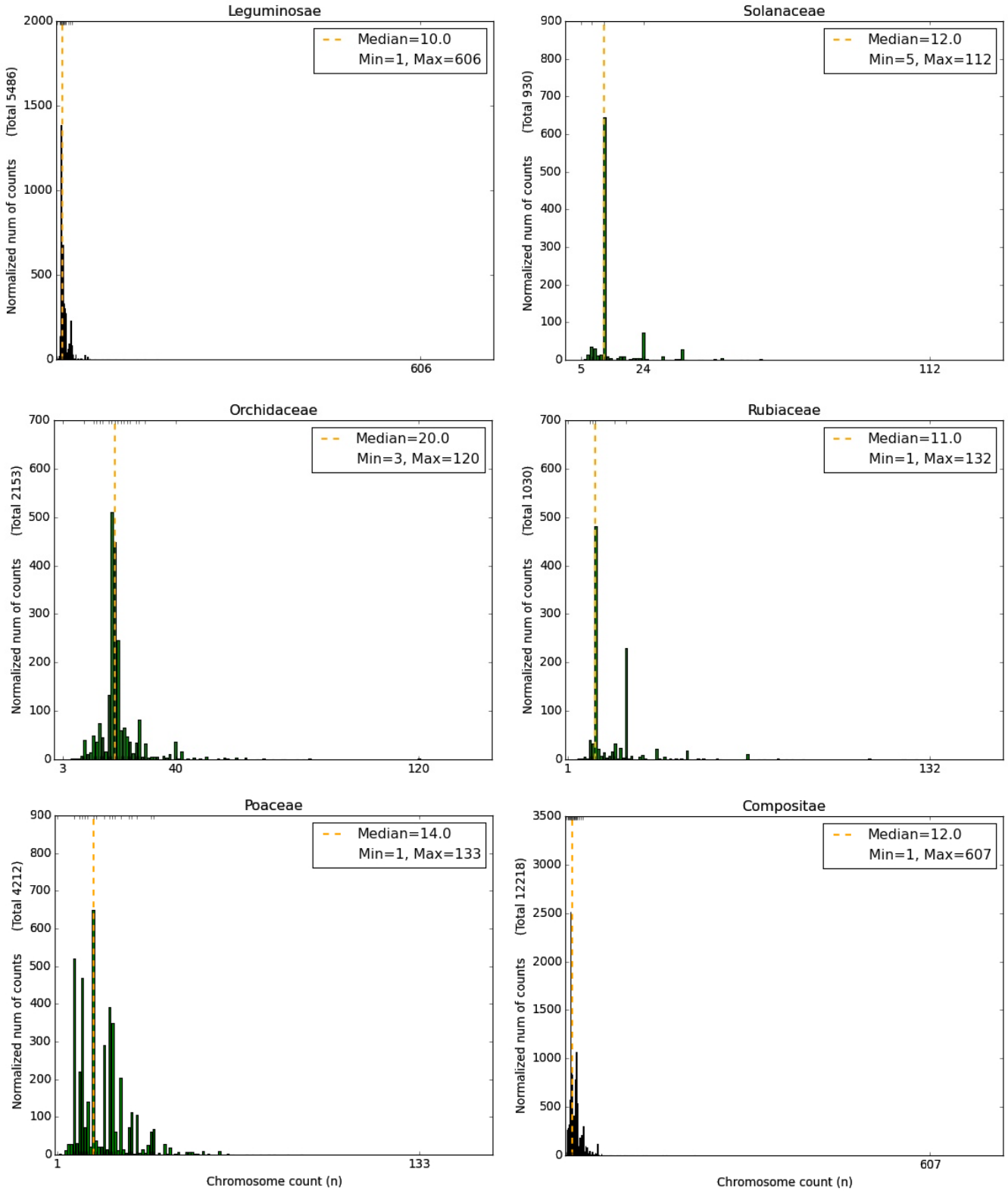


Figure 5.3 Chromosome count distributions for the five largest angiosperm families according to APG III (2009), Asteraceae; Orchidaceae; Fabaceae; Rubiaceae; Poaceae; plus Solanaceae. Statistics downloaded from the Chromosome Counts Database (Rice *et al.*, 2014; http://ccdb.tau.ac.il/).

Ecology of Nicotiana *section* Suaveolentes

Further work is sorely needed on the ecology and pollination biology of *Nicotiana* section *Suaveolentes*. Almost nothing is known about the pollinators of these plants, but most species appear to be selfing to a high degree. Investigating the mechanism of selfing would also be an important contribution to understanding the reproductive biology of these species, although it probably occurs simply through close proximity of anthers and stigma – as found in self-compatible populations of *N. benthamiana* (Bally *et al*., 2015).

For many species of *Suaveolentes* self-compatibility and self-fertilisation are clearly linked with a rapid (annual) life cycle, which is an adaptation to the Eremaean zone of central Australia. During fieldwork it was noted that several species have particular ecological requirements, either in terms of substrate or relative shade. Some of these details have been noted in the descriptions for species and particularly in the Flora of Australia (Purdie *et al*., 1982). For example, *N. velutina* is mostly, but not exclusively found on red sand dunes; *N. truncata* is found only on gypseous friable cracking soils (Symon, 1981) off the gibber plain but not on the plain itself; *N. burbidgeae* occurs at the base of cutaway clay ridges and *N. faucicola* is found on the sides of deep gorges in the Flinders Ranges (Figure 5.4). Thus in addition to distinct morphology amongst species, there are also distinct ecologies that are worthy of further research, particularly in order to find out what has driven the diversification of this group.

Figure 5.4 Habitat of Australian members of *Nicotiana* section *Suaveolentes*. A – *N. truncata*; B – *N. burbidgeae*; C – *N. velutina*; D – *N. faucicola*.

Tales of plastome-nuclear discordance

Incongruence between plastid and ribosomal DNA is something that has pervaded the systematics literature for many years. In the case of *Nicotiana* section *Suaveolentes* there is a large degree of congruence between plastomes and the rDNA cistron (Chapter 4). However, there are some notable exceptions, and several cases in which the rDNA tree groups with strong support populations of a species that were otherwise widely dispersed in the plastome tree. Out of the possible explanations for this, the two most likely are plastome introgression and incomplete lineage sorting (ILS)/retention of ancestral polymorphism. The relative differences found in the trees suggest ILS as more likely than recent hybridisation. This is most prominent in the clade that contains a number of species with the same chromosome number of n = 16, so a pattern of introgression could be possible via sexual means in this clade, though misidentification of some taxa is also possible. The potential of plastid introgression via non-sexual means has been reported, for instance recently it has been shown to be possibly by vegetative grafting (Stegemann *et al.*, 2012) in *Nicotiana*. The problem with this is that it seems extremely unlikely to occur for annual, herbaceous species in nature, although it definitely highlights the potential for horizontal transfer of the plastome without interbreeding of species.

In *Suaveolentes* the morphological and ecological boundaries between some of the taxa involved in these instances of plastid/nuclear tree discordance are clearly distinct. Thus, it seems unlikely that rampant hybridisation and introgression are occurring as this would more often than not blur these boundaries between taxa. For example, *N. goodspeedii*, *N. velutina* and *N. faucicola* are clearly different morphologically but entirely mixed up in the velutina clade of the plastome tree. The narrow range endemic *N. truncata* is phenomenally distinct morphologically amongst *Suaveolentes* species, with succulent glabrous leaves in a rosette, glabrous inflorescences and stout but

substantial flowers with truncate calyces. Despite this, it occurs in two entirely separate clades in the plastome tree, and yet forms a highly supported clade in the rDNA tree. As a whole, these results strongly suggest that ancestral polymorphism and ILS are more likely than introgression between species of section *Suaveolentes*. Clearly this discordance requires further investigation.

Future prospects of genome skimming for phylogenomics

Genome skimming is an attractive high-throughput sequencing approach due to its simplistic lab protocol and relative affordability (Dodsworth, 2015) and has been used increasingly in plants to answer phylogenetic questions at various levels (Bock *et al*., 2014; Kane *et al*., 2012; Malé *et al*., 2014; McPherson *et al*., 2013; Ruhsam *et al*., 2015) including those previously thought unanswerable without HTS technologies (Knapp, 2014). Variability of plastid DNA in genome skims is something that requires further analysis and may be more related to the age and developmental stage of material (Rowan and Bendich, 2009) than genome size. Although it can be argued that the plastome represents only one locus (a non-recombining region), and the associated issue of gene conversion in rDNA does the same for this region, the relative ease of genome skimming and the possibilities that the plastome alone provide are still arguably worth adopting this approach. For systematists, this signifies orders of magnitude more sequence data (~150 kb relative to typically less than 10 kb in traditional Sanger sequencing studies), but it can be amazingly invariant in recently radiated groups of species (e.g. *Diospyros* on New Caledonia, for which complete plastome sequencing provided only 350 variable positions among 25 clearly morphologically and ecologically distinct species; Paun *et al*., 2015). In this thesis I present the results of genome skimming in a recent radiation of Solanaceae, with over 100 samples and demonstrate the feasibility of this methodology and the insights it brings to a poorly known group within the genus *Nicotiana*.

Future directions for Nicotiana *section* Suaveolentes *research*

Further insight could be gained from large-scale analyses of genes in section *Suaveolentes*, either through a transcriptomic or target enrichment approach, and indeed the latter are a popular focus of phylogenomics research groups at the current time of writing. RADSeq is another approach currently under consideration, that has been shown to be extremely useful in resolving recent radiations. Certain caveats should be mentioned regarding target enrichment because it involves the complex design of bait sequences and associated cost of producing the baits. This is bioinformatically and economically still out of the reach of many systematics labs, and thus a more simplistic genome skimming approaches may be more appropriate.

Nonetheless, a large gene dataset (perhaps obtained by using RADSeq) would certainly add to analyses of selection and adaptation, and would likely add support to the phylogenetic hypothesis for *Nicotiana* section *Suaveolentes*. It should be noted that in spite of this prospect the lineage sorting issue found in the current study will probably be magnified in analyses of hundreds or thousands of genes – i.e. the proportion of loci that are incongruent and reflect ILS may remain relatively constant as the number of loci are increased. The occurrence of rampant ILS is frequently becoming apparent as recent radiations are probed with ever-greater genomic sampling, (e.g. Carbone *et al.*, 2014; Kutschera *et al.*, 2014; Heyduk *et al.*, 2015), and even with sophisticated coalescent methods the species tree is sometimes difficult to infer. Speciation may involve only a handful of genes under strong selection, leaving the rest of the genome a hodgepodge of incongruent gene trees as a result of old ILS and recent introgression. In these cases, the uncertainty in phylogenomic inference is a true reflection of the underlying biology of rapid speciation and therefore an interesting result in itself.

References

Abraham A, Ninan CA. 1954. The chromosomes of *Ophioglossum reticulatum* L. *Current Science* **23:** 213–214.

Aoki S, Ito M. 2000. Molecular phylogeny of *Nicotiana* (Solanaceae) based on the nucleotide sequence of the *matK* gene. *Plant Biology* **2:** 316-324.

Aflitos S, Schijlen E, de Jong H, de Ridder D, Smit S, Finkers R, Wang J, Zhang G, Li N, Mao L, Bakker F, Dirks R, Breit T, Gravendeel B, Huits H, Struss D, Swanson-Wagner R, van Leeuwen H, van Ham RCHJ, Fito L, Guignier L, Sevilla M, Ellul P, Ganko E, Kapur A, Reclus E, de Geus B, de van Geest H, te Lintel Hekkert B, van Haarst J, Smits L, Koops A, Sanchez-Perez G, van Heusden AW, Visser R, Quan Z, Min J, Liao L, Wang X, Wang G, Yue Z, Yang X, Xu N, Schranz E, Smets E, Vos R, Rauwerda J, Ursem R, Schuit C, Kerns M, van den Berg J, Vriezen W, Janssen A, Datema E, Jahrman T, Moquet F, Bonnet J, Peters S. 2014. Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole genome sequencing. *The Plant Journal* **80:** 136–148.

Ambrozova K, Mandakova T, Bures P, Neumann P, Leitch IJ, Koblizkova A, Macas J, Lysak MA. 2011. Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria* lilies. *Annals of Botany* **107:** 255–268.

Angiosperm Phylogeny Group. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* **161**: 105–121.

Avise JC, Ball RM. 1990. Principles of genealogical concordance in species concepts and biological taxonomy. *Oxford Surveys in Evolutionary Biology* **7:** 45-67.

Bai C, Alverson WS, Follansbee A, Waller DM. 2012. New reports of nuclear DNA content for 407 vascular plant taxa from the United States. *Annals of Botany* **110:** 1623–1629.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3:** e3376.

Bally J, Nakasugi K, Jia F, Jung H, Ho SYW, Wong M, Paul CM, Naim F, Wood CC, Crowhurst RN, Hellens RP, Dale JL, Waterhouse PM. 2015. The extremophile *Nicotiana benthamiana* has traded viral defence for early vigour. *Nature Plants* **1:** article number 15165.

Barrett CF, Davis JI, Leebens-Mack J, Conran JG, Stevenson DW. 2013. Plastid genomes and deep relationships among the commelinid monocot angiosperms. *Cladistics* **29:** 65–87.

Blanc G, Wolfe KH. 2004. Widespread palaeopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell* **16:** 1667–1678.

Bock DG, Kane NC, Ebert DP, Rieseberg LH. 2014. Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. *New Phytologist* **201:** 1021–1030.

Bombarely A, Rosli HG, Vrebalov J, Moffett P, Mueller LA, Martin GB. 2012. A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. *Molecular Plant-Microbe Interactions* **25:** 1523-1530.

Bowers JE, Chapman BA, Rong JK, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422:** 433–438.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* **10:** e1003537.

Bouckaert R, Heled J. 2014. DensiTree2: Seeing trees through the forest. *bioRxiv* DOI: 10.1101/012401.

Brookfield JFY. 2005. The ecology of the genome – mobile DNA elements and their hosts. *Nature Reviews Genetics* **6:** 128–136.

Bybee SM, Bracken-Grissom H, Haynes BD, Hermansen RA, Byers RL, Clement MJ, Udall JA, Wilcox ER, Crandall KA. 2011. Targeted amplicon sequencing (TAS): A scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biology and Evolution* **3:** 1312–1323.

Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer TJ, Herrero J, Roos C, Aken B, *et al*. 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513:** 195–201.

Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, Sikora M, Rasmussen M, Gravel S, Guillén S, Nekhrizov G, Leshtakov K, Dimitrova D, Theodossiev N, Pettener D, Luiselli D, Sandoval K, Morena-Estrada A, Li Y, Wang J, Gilbert TP, Willerslev E, Greenleaf WJ, Bustamante CD. 2013. Pulling out the 1%: Whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *American Journal of Human Genetics* **93:** 852–864.

CBOL Working Plant Group. 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* **106:** 12794–12797.

Chase MW, Hills HH. 1991. Silica gel: An ideal material for field preservation of leaf samples for DNA studies. *Taxon* **2:** 215-220.

Chase MW, Knapp S, Cox AV, Clarkson JJ, Butsko Y, Joseph J, Savolainen V, Parokonny AS. 2003. Molecular systematics, GISH and the origin of hybrid taxa in *Nicotiana* (Solanaceae). *Annals of Botany* **92:** 107–127.

Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madriñan S, Petersen G, Seberg O, Jørgsensen T, Cameron KM, Carine M, Pederson N, Hedderson TAJ, Conrad F, Salazar GA, Richardson JE, Hollingsworth ML, Barraclough TG, Kelly LJ, Wilkinson M. 2007. A proposal for a standardised protocol to barcode all land plants. *Taxon* **56:** 295-299.

Chester M, Gallagher JP, Vaughan Symonds V, Veruska Cruz da Silva A, Mavrodiev EV, Leitch AR, Soltis PS, Soltis DE. 2012. Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus*. *Proceedings of the National Academy of Sciences* **109:** 1176–1181.

Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)* **99:** 45-56.

Clarkson JJ, Knapp S, Garcia VF, Olmstead RG, Leitch AR, Chase MW. 2004. Phylogenetic relationships in *Nicotiana* (Solanaceae) inferred from multiple plastid DNA regions. *Molecular Phylogenetics and Evolution* **33:** 75–90.

Clarkson JJ, Lim KY, Kovarik A, Chase MW, Knapp S, Leitch AR. 2005. Long-term genome diploidisation in allopolyploid *Nicotiana* section *Repandae* (Solanaceae). *New Phytologist* **168:** 241–252.

Clarkson JJ. 2007. Evolutionary relationships in *Nicotiana* (Solanaceae). PhD thesis: Queen Mary, University of London.

Clarkson JJ, Kelly LJ, Leitch AR, Knapp S, Chase MW. 2010. Nuclear glutamine synthetase evolution in *Nicotiana*: Phylogenetics and the origins of allotetraploid and homoploid (diploid) hybrids. *Molecular Phylogenetics and Evolution* **55:** 99–112.

Conant GC, Birchler JA, Pires CP. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion in Plant Biology* **19:** 91–98.

Cranston KA, Hurwitz B, Ware D, Stein L, Wing RA. 2009. Species trees from highly incongruent gene trees in rice. *Systematic Biology* **58:** 489-500.

Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J. 2012. Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany* **99:** 291–311.

Cummings MP, Neel MC, Shaw KL. 2008. A genealogical approach to quantifying lineage divergence. *Evolution* **62:** 2411-2422.

Day PD, Berger M, Hill L, Fay MF, Leitch AR, Leitch IJ, Kelly LJ. 2014. Evolutionary relationships in the medicinally important genus *Fritillaria* L. (Liliaceae). *Molecular Phylogenetics and Evolution* **80:** 11–19.

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single copy families in flowering plants. *Proceedings of the National Academy of Sciences* **110:** 2898–2903.

Dodsworth S. 2015. Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science* **20:** 525-527.

Dodsworth S, Chase MW, Kelly LJ, Leitch IJ, Macas J, Novák P, Piednoël M, Weiss-Schneeweiss H, Leitch AR. 2015a. Genomic repeat abundances contain phylogenetic signal. *Systematic Biology* **64:** 112–126.

Dodsworth S, Chase MW, Särkinen T, Knapp S, Leitch AR. 2015b. Using genomic repeats for phylogenomics: a case study in wild tomatoes (*Solanum* section *Lycopersicon*: Solanaceae). *Biological Journal of the Linnean Society* DOI: 10.1111/bij.12612.

Dodsworth S, Leitch AR, Leitch IJ. 2015c. Genome size diversity in angiosperms and its influence on gene space. *Current Opinion in Genetics & Development* **35:** DOI 10.1016/j.gde.2015.10.006.

Dodsworth S, Chase MW, Leitch AR. 2015d. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society* DOI: 10.1111/boj.12357.

Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4:** e88.

Edwards SV, Fertil B, Giron A, Deschavanne PJ. 2002. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Systematic Biology* **51:** 599–613.

Fedoroff NV. 2012. Transposable elements, epigenetics, and genome evolution. *Science* **338:** 758–767.

Felsenstein J. 1989. PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics* **5:** 164–166.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. *New Phytologist* **183:** 557–564.

Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA, Mummenhoff K. 2011. Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends in Plant Science* **16:** 108–116.

Gillett CPDT, Crampton-Platt A, Timmermans MJTN, Jordal BH, Emerson BC, Vogler AP. 2014. Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution* **31:** 2223–2237.

Glenn TC. 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11:** 759-769.

Glick L, Mayrose I. 2014. ChromEvol: Assessing the pattern of chromosome number evolution and the inference of polyploidy along a phylogeny. *Molecular Biology and Evolution* **31:** 1914-1922.

Goloboff PA, Farris JS, Nixon K. 2003a. Tree analysis using new technology [Internet]. Program and documentation. Distributed by the authors. Available from: http://www.zmuc.dk/public/phylogeny.

Goloboff PA, Farris JS, Kallersjo M, Oxelman B, Ramirez MJ, Szumik CA. 2003b. Improvements to resampling measures of group support. *Cladistics* **19:** 324–332.

Goloboff PA, Mattoni CI, Quinteros AS. 2006. Continuous characters analyzed as such. *Cladistics* **22:** 589–601.

Goloboff PA, Farris JS, Nixon KC. 2008. TNT, a free program for phylogenetic analysis. *Cladistics* **24:** 774–786.

Goodspeed, TH. 1954. The genus *Nicotiana*. Chronica Botanica Company: Waltham, Mass.

Gorelick R, Olson K. 2013. Polyploidy is genetic hence may cause non-adaptive radiations, whereas pseudopolyploidy is genomic hence may cause adaptive non-radiations. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **5:** 286–294.

Grandillo S, Chetelat R, Knapp S, Spooner D, Peralta E, Cammareri M, Perez O, Termolino P, Tripodi P, Chuisano ML, Ercolano MR, Frusciante L, Monti L, Pignone D. 2011. *Solanum* section *Lycopersicon*. In: Kole C, ed. Wild crop relatives: genomic and breeding resources. Berlin & Heidelberg: Springer Verlag, 129–215.

Greilhuber J, Leitch IJ. 2013. Genome size and the phenotype. In: Leitch IJ, Greilhuber J, Doležel J, Wendel J, eds. Plant genome diversity, Vol. 2. Physical structure, behavior and evolution of plant genomes. London: Springer.

Guignard MS, Nichols RA, Knell RJ, Macdonald A, Romila C-A, Trimmer M, Leitch IJ, Leitch AR. 2015. Genome size and ploidy influence angiosperm species biomass under nitrogen and phosphorus limitation. *New Phytologist*, submitted.

Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen E, Sonet G, Nagy ZT, Lenglet G, Mayer F, Savolainen V. 2013. Next-generation museomics disentangles one of the largest primate radiations. *Systematic Biology* **62:** 539–554.

Hansen CN, Heslop-Harrison JS. 2004. Sequences and phylogenies of plant pararetroviruses, viruses and transposable elements. *Advances in Botanical Research* **41:** 165–193.

Haran J, Timmermans MJTN, Vogler AP. 2013. Mitogenome sequences stabilize the phylogenetics of weevils (Curculionoidea) and establish the monophyly of larval ectophagy. *Molecular Phylogenetics and Evolution* **67:** 156–166.

Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B* **270:** 313–321.

Heled J, Drummond AJ. 2010. Bayesian inference of species tree from multilocus data. *Molecular Biology and Evolution* **27:** 570-580.

Heled J, Drummond AJ. 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Systematic Biology* **61:** 138-149.

Heyduk K, Trapnell DW, Barrett CF, Leebens-Mack J. 2015. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biological Journal of the Linnean Society* DOI: 10.1111/bij.12551.

Horton P. 1981. A taxonomic revision of *Nicotiana* (Solanaceae) in Australia. *Journal of the Adelaide Botanic Gardens* **3:** 1-56.

Husband BC. 2000. Constraints on polyploid evolution: a test of the minority cytotype exclusion principle. *Proceedings of the Royal Society B* **267:** 217–233.

Hunt T, Bergsten J, Levkanicova Z, Papadopoulou A, St. John O, Wild R, Hammond PR, Ahrens D, Balke M, Caterino MS, Gómez-Zurita J, Ribera I, Barraclough TG, Bocakova M, Bocak L, Vogler AP. 2007. A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science* **21:** 1913-1916.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23:** 254–267.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderball AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, de Pamphilis CW. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473:** 97–100.

Jurka J, Bao W, Kojima KK. 2011. Families of transposable elements, population structure and the origin of species. *Biology Direct* **6:** 44.

Jurka J, Bao W, Kojima KK, Kohany O, Yurka MG. 2012. Distinct groups of repetitive families preserved in mammals correspond to different periods of regulatory innovations in vertebrates. *Biology Direct* **7:** 36.

Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JMM, Cronk Q. 2012. Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany* **99:** 320–329.

Katoh K, Standley DM. 2014. MAFFT: iterative refinement and additional methods. *Methods in Molecular Biology* **1079:** 131-146.

Kayal E, Roure B, Philippe H, Collins AG, Lavrov DV. 2013. Cnidarian phylogenetic relationships as revealed by mitogenomics. *BMC Evolutionary Biology* **13:** 5.

Kejnovsky E, Leitch IJ, Leitch AR. 2009. Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends in Ecology & Evolution* **24:** 572–582.

Kelly LJ, Leitch AR, Clarkson JJ, Hunter RB, Knapp S, Chase MW. 2010. Intragenic recombination events and evidence for hybrid speciation in *Nicotiana* (Solanaceae). *Molecular Biology and Evolution* **27:** 781–799.

Kelly LJ, Leitch IJ. 2011. Exploring giant plant genomes with next-generation sequencing technology. *Chromosome Research* **19:** 939–953.

Kelly LJ, Leitch AR, Fay MF, Renny-Byfield S, Pellicer J, Macas J, Leitch IJ. 2012. Why size really matters when sequencing plant genomes. *Plant Ecology & Diversity* **5:** 415–425.

Kelly LJ, Leitch AR, Clarkson JJ, Knapp S, Chase MW. 2013. Reconstructing the complex evolutionary origin of wild allopolyploid tobaccos (*Nicotiana* section *Suaveolentes*). *Evolution* **67:** 80–94.

Knapp S, Chase MW, Clarkson JJ. 2004. Nomenclatural changes and a new sectional classification in *Nicotiana* (Solanaceae). *Taxon* **53:** 73-82.

Knapp S. 2014. Why is a raven like a writing desk? Origins of the sunflower that is neither an artichoke nor from Jerusalem. *New Phytologist* **201:** 710-711.

Knoop V. 2004. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. *Current Genetics* **46:** 123-139.

Kovarik A, Renny-Byfield S, Grandbastien M-A, Leitch AR. 2012. Evolutionary implications of genome and karyotype restructuring in *Nicotiana tabacum* L. In: Soltis P.S., Soltis D.E., editors. Polyploidy and Genome Evolution. London: Springer.

Kutschera VE, Bidon T, Hailer F, Rodi JL, Fain SR, Janke A. 2014. Bears in a forest of gene trees: Phylogenetic inference is complicated by incomplete lineage sorting and gene flow. *Molecular Biology and Evolution* **31:** 2004-2017.

Ladiges PY, Marks CE, Nelson G. 2011. Biogeography of *Nicotiana* section *Suaveolentes* (Solanaceae) reveals geographical tracks in arid Australia. *Journal of Biogeography* **38:** 2066-2077.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* **21:** 1095-1109.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25:** 2286-2288.

Leitch IJ, Bennett MD. 2004. Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society* **82:** 651–663.

Leitch AR, Leitch IJ. 2008. Genomic plasticity and the diversity of polyploid plants. *Science* **320:** 481–483.

Leitch IJ, Hanson L, Lim KY, Kovarik A, Chase MW, Clarkson JJ, Leitch AR. 2008. The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). *Annals of Botany* **101:** 805-814.

Leitch AR, Leitch IJ. 2012. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytologist* **194:** 629–646.

Levin DA. 1975. Minority cytotype exclusion in local plant populations. *Taxon* **24:** 35–43.

Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S. 2015. Plant DNA barcoding: from gene to genome. *Biological Reviews* **90:** 157–166.

Lim KY, Matyasek R, Lichtenstein CP, Leitch AR. 2000. Molecular cytogenetic analyses and phylogenetic studies in *Nicotiana* section *Tomentosae*. *Chromosoma* **109:** 245–258.

Lim KY, Kovarik A, Matyasek R, Chase MW, Knapp S, McCarthy E, Clarkson JJ, Leitch AR. 2006. Comparative genomics and repetitive sequence divergence in the species of diploid *Nicotiana* section *Alatae*. *The Plant Journal* **48:** 907–919.

Lim KY, Kovarik A, Matyasek R, Chase MW, Clarkson JJ, Grandbastien MA, Leitch AR. 2007. Sequence of events leading to near-complete genome turnover in allopolyploid *Nicotiana* within five million years. *New Phytologist* **175:** 756-763.

Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, Huang Z, Li J, Zhang C, Wang T, Zhang Y, Wang A, Zhang Y, Lin K, Li C, Xiong G, Xue Y, Mazzucato A, Causse M, Fei Z, Giovannoni JJ, Chetelat RT, Zamir D, Städler T, Li J, Ye Z, Du Y, Huang S. 2014. Genomic analyses provide insights into the history of tomato breeding. *Nature Genetics* **46:** 1220–1226.

Macas J, Neumann P, Navratilova A. 2007. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterisation using 454 sequencing and comparison to soybean and *Medicago trunculata*. *BMC Genomics* **8:** 427.

Maddison WP. 1997. Gene trees in species trees. *Systematic Biology* **46:** 523-536.

Malé PJG, Bardon L, Besnard G, Coissac E, Delsuc F, Engel J, Lhuillier E, Scotti-Saintagne C, Tinaut A, Chave J. 2014. Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Molecular Ecology Resources* **14:** 966–975.

Mándaková T, Mummenhoff K, Al-Shehbaz IA, Mucina L, Mühlhausen A, Lysak MA. 2012. Whole-genome triplication and species radiation in the southern African tribe Heliophileae (Brassicaceae). *Taxon* **61:** 989–1000.

Marks CE. 2010a. The evolution of *Nicotiana* section *Suaveolentes*. PhD thesis: University of Melbourne.

Marks CE. 2010b. Definition of South Pacific taxa of *Nicotiana* section *Suaveolentes* (Solanaceae). *Muelleria* **28:** 74-84.

Marks CE, Newbigin E, Ladiges PY. 2011a. Comparative morphology and phylogeny of *Nicotiana* section *Suaveolentes* (Solanaceae) in Australia and the South Pacific. *Australian Systematic Botany* **24:** 61-86.

Marks CE, Ladiges PY, Newbigin E. 2011b. Karyotypic variation in *Nicotiana* section *Suaveolentes*. *Genetic Resources and Crop Evolution* **58:** 797-803.

Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, Otto SP. 2011. Recently formed polyploid plants diversify at lower rates. *Science* **333:** 1257.

McCarthy EW, Arnold SEJ, Chittka L, Le Comber SC, Verity R, Dodsworth S, Knapp S, Kelly LJ, Chase MW, Baldwin IT, Kovarik A, Mhiri C, Taylor L, Leitch AR. 2015. The effect of polyploidy and hybridization on the evolution of floral colour in *Nicotiana* (Solanaceae).

McClintock B. 1984. The significance of responses of the genome to challenge. *Science* **226:** 792–801.

Meudt HM, Rojas-Andrés BM, Prebble JM, Low E, Garnock-Jones PJ, Albach DC. 2015. Is genome downsizing associated with diversification in polyploid lineages of *Veronica*? *Botanical Journal of the Linnean Society* **178:** 243–266.

McPherson H, van der Merwe M, Delaney SK, Edwards MA, Henry RJ, McIntosh E, Rymer PD, Milner ML, Siow J, Rossetto M. 2013. Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. *BMC Ecology* **13:** 8.

Meier R, Zhang G, Ali F. 2008. The use of mean instead of smallest interspecific distances exaggerates the size of the 'barcoding gap' and leads to misidentification. *Systematic Biology* **57:** 809–813.

Meyer CP, Paulay G. 2005. DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology* **3:** e422.

Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M. 2007. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research* **35:** e97.

Miller M, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES science gateway for inference of large phylogenetic trees. In Gateway Computing Environments Workshop (GCE), 1-8.

Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* **14:** 193-202.

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. 1986. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology* **51:** 263-273.

Naim F, Nakasugi K, Crowhurst RN, Hilario E, Zwart AB, Hellens RP, Taylor JM, Waterhouse PM, Wood CC. 2012. Advanced engineering of lipid metabolism in *Nicotiana benthamiana* using a draft genome and the V2 viral silencing-suppressor protein. *PLoS One* **7:** e52717.

Neiman M, Kay AD, Krist AC. 2013. Can resource costs of polyploidy provide an advantage to sex? *Heredity* **110:** 152–159.

Neumann P, Navratilova A, Schroeder-Reiter E, Koblizkova A, Steinbauerova V, Chocholova E, Novák P, Wanner G, Macas J. 2012. Stretching the rules: Monocentric chromosomes with multiple centromere domains. *PLoS Genetics* **8:** e1002777.

Njuguna W, Liston A, Cronn R, Ashman TL, Bassil N. 2013. Insights into phylogeny, sex function and age of *Fragaria* based on whole chloroplast genome sequencing. *Molecular Phylogenetics and Evolution* **66:** 17–29.

Nichols R. 2001. Gene trees and species trees are not the same. *Trends in Ecology & Evolution* **16:** 358-364.

Novák P, Neumann P, Macas J. 2010. Graph-based clustering and characterisation of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* **11:** 378.

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J. 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterisation of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29:** 792–793.

Obbard DJ, Maclennan J, Kim KW, Rambaut A, O'Grady PM, Jiggins FM. 2012. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Molecular Biology and Evolution* **29:** 3459–3473.

Oliver M, Petrov D, Ackerly D, Falkowski P, Schofield OM. 2007. The mode and tempo of genome size evolution in eukaryotes. *Genome Research* **17:** 594–601.

Otto S. 2007. The evolutionary consequences of polyploidy. *Cell* **131:** 452–462.

Pagan HJT, Macas J, Novák P, McCulloch ES, Stevens RD, Ray DA. 2012. Survey sequencing reveals elevated DNA transposon activity, novel elements, and variation in repetitive landscapes among bats. *Genome Biology and Evolution* **4:** 575–585.

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* **401:** 877-884.

Palmer JD. 1992. Comparison of chloroplast and mitochondrial genome evolution in plants. In R.G. Hermann (ed.) Cell organelles. Springer: Berlin Heidelberg, New York.

Parisod C, Mhiri C, Lim KY, Clarkson JJ, Chase MW, Leitch AR, Grandbastien M-A. 2012. Differential dynamics of transposable elements during long-term

diploidisation of *Nicotiana* section *Repandae* (Solanaceae) allopolyploid genomes. *PLoS One* **7:** e50352.s.

Park JM, Manen JF, Colwell AE, Schneeweiss GM. 2008. A plastid gene phylogeny of the non-photosynthetic parasitic *Orobanche* (Orobanchaceae) and related genera. *Journal of Plant Research* **121:** 365–376.

Paun O, Turner B, Trucchi E, Munzinger J, Chase MW, Samuel R. 2015. Processes driving the adaptive radiation of a tropical tree (*Diospyros*, Ebenacaee) in New Caledonia, a biodiversity hotspot. *Systematic Biology* doi: 10.1093/sysbio/syv076.

Pellicer J, Fay MF, Leitch IJ. 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society* **164:** 10–15.

Pellicer J, Leitch IJ. 2014. The application of flow cytometry for estimating genome size and ploidy level in plants. In P. Besse (ed.) Molecular Plant Taxonomy. Humana Press, 279-307.

Peralta IE, Knapp S, Spooner DM. 2005. New species of wild tomatoes (*Solanum* section *Lycopersicon*: Solanaceae) from Northern Peru. *Systematic Botany* **30:** 424–434.

Peralta IE, Spooner DM, Knapp S. 2008. Taxonomy of wild tomatoes and their relatives (*Solanum* sect. *Lycopersoides*, sect. *Juglandifolia*, sect. *Lycpersicon*; Solanaceae). *Systematic Botany Monographs* **84:** 1–186.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS One* **7:** e37135.

Piednoël M, Aberer AJ, Schneeweiss GM, Macas J, Novák P, Gundlach H, Temsch EM, Renner SS. 2012. Next-generation sequencing reveals the impact of repetitive DNA across phylogenetically closely-related genomes of Orobanchaceae. *Molecular Biology and Evolution* **29:** 3601–3611.

Piednoël M, Carrete-Vega G, Renner SS. 2013. Characterization of the LTR retrotransposon repertoire of a plant clade of six diploid and one tetraploid species. *The Plant Journal* **75:** 699–709.

Pillon Y, Johansen J, Sakishima T, Chamala S, Barbazuk WB, Roalson EH, Price DK, Stacy EA. 2013. Potential use of low-copy nuclear genes in DNA barcoding: a comparison with plastid genes in two Hawaiian plant radiations. *BMC Evolutionary Biology* **13:** 35.

Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Research* **13:** 145–158.

Purdie RW, Symon DE, Haegi L. 1982. Solanaceae. In Flora of Australia. Australian Government Publishing Service: Canberra.

Renny-Byfield S, Chester M, Kovarik A, LeComber SC, Grandbastien M-A, Deloger M, Nichols RA, Macas J, Novak P, Chase MW, Leitch AR. 2011. Next generation sequencing reveals genome downsizing in allotetraploid Nicotiana tabacum, predominantly through the elimination of paternally derived repetitive DNAs. *Molecular Biology and Evolution* **28:** 2843–2854.

Renny-Byfield S, Kovarik A, Chester M, Nichols RA, Macas J, Novák P, Leitch AR. 2012. Independent, rapid and targeted loss of highly repetitive DNA in natural and synthetic allopolyploids of *Nicotiana tabacum*. *PLoS One* **7:** e36963.

Renny-Byfield S, Kovarik A, Kelly LJ, Macas J, Novak P, Chase MW, Nichols RA, Pancholi MR, Grandbastien M-A, Leitch AR. 2013. Diploidisation and genome size change in allopolyploids is associated with differential dynamics of low and high copy sequences. *The Plant Journal* **74:** 829–839.

Renny-Byfield S, Wendel JF. 2014. Doubling down on genomes: polyploidy and crop plants. *American Journal of Botany* **101:** 1711-1725.

Rice A, Glick L, Abadi S, Einhorn M, Kopelman NM, Salman-Minkov A, Mayzel J, Chay O, Mayrose I. 2014. The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. *New Phytologist* **206:** 19-26.

Rix M. 2001. *Fritillaria*, a revised classification. The *Fritillaria* Group of the Alpine Garden Society: UK.

Rodriguez F, Wu F, Ané C, Tanksley S, Spooner DM. 2009. Do potatoes and tomatoes have a single evolutionary history, and what proportion of the genome supports this history? *BMC Evolutionary Biology* **9:** 191.

Rønsted N, Law S, Thornton H, Fay MF, Chase MW. 2005. Molecular phylogenetic evidence for the monophyly of *Fritillaria* and *Lilium* (Liliaceae; Liliales) and the infrageneric classification of *Fritillaria*. *Molecular Phylogenetics and Evolution* **35:** 509–527.

Rowan BA, Bendich AJ. 2009. The loss of DNA from chloroplasts as leaves mature: fact or artefact? *Journal of Experimental Botany* **60:** 3005–3010.

Rubin BER, Ree RH, Moreau CS. 2012. Inferring phylogenies from RAD sequence data. *PLoS One* **7:** e33394.

Ruhsam M, Rai HS, Mathews S, Ross TG, Graham SW, Raubeson LA, Mei W, Thomas PI, Gardner MF, Ennos RA, Hollingsworth PM. 2015. Does complete plastid genome sequencing improve species discrimination and phylogenetic resolution in *Araucaria*? *Molecular Ecology Resources* **15:** 1067-1078.

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239:** 487-491.

Sakalidis ML, Hardy GESJ, Burgess TI. 2011. Use of the genealogical sorting index (GSI) to dileneate species boundaries in the *Neofusicoccum parvum-Neofusicoccum ribis* species complex. *Molecular Phylogenetics and Evolution* **60:** 333-344.

Särkiinen T, Staats M, Richardson JE, Cowan RS, Bakker FT. 2012. How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLoS One* **7:** e43808.

Särkiinen T, Bohs L, Olmstead RG, Knapp S. 2013. A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. *BMC Evolutionary Biology* **13:** 214.

Schaefer H, Hechenleitner P, Santos-Guerra A, Menezes de Sequeira M, Pennington RT, Kenicer G, Carine MA. 2012. Systematics, biogeography, and character evolution of the legume tribe Fabeae with special focus on the middle-Atlantic island lineages. *BMC Evolutionary Biology* **12:** 250.

Schmidt-Lebuhn AN, de Vos JM, Keller B, Conti E. 2012. Phylogenetic analysis of *Primula* section *Primula* reveals rampant non-monophyly among morphologically distinct species. *Molecular Phylogenetics and Evolution* **65:** 23-34.

Schneeweiss GM, Colwell AE, Park JM, Jang CG, Stuessy TF. 2004. Phylogeny of holoparasitic *Orobanche* (Orobanchaceae) inferred from nuclear ITS-sequences. *Molecular Phylogenetics and Evolution* **30:** 465–478.

Schranz ME, Mohammadin S, Edger PE. 2012. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. *Current Opinion in Plant Biology* **15:** 147–153.

Scowcroft WR. 1979. Nucleotide polymorphism in chloroplast DNA of *Nicotiana debneyi*. *Theoretical and Applied Genetics* **55:** 133-137.

Seetharam AS, Stuart GW. 2013. Whole genome phylogeny for 21 *Drosophila* species using predicted 2b-RAD fragments. *PeerJ* **1:** e226.

Sierro N, Battey JND, Ouadi S, Bovet L, Goepfert S, Bakaher N, Peitsch MC, Ivanov NV. 2013. Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. *Genome Biology* **14:** R60.

Sierro N, Battey JND, Ouadi S, Bakaher N, Bovet L, Willig A, Goepfert S, Peitsch MC, Ivanov NV. 2014. The tobacco genome sequence and its comparison with those of tomato and potato. *Nature Communications* **5:** article number 3833.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* **321:** 674-679.

Smith UE, Hendricks JR. 2013. Geometric morphometric character suites as phylogenetic data: Extracting phylogenetic signal from gastropod shells. *Systematic Biology* **62:** 366–385.

Šmarda P, Hejcman M, Brezinová A, Horová L, Steigerová H, Zedek F, Bureš P, Hejcmanová P, Schellberg J. 2013. Effect of phosphorus availability on the selection of species with different ploidy levels and genome sizes in a long-term grassland fertilization experiment. *New Phytologist* **200:** 911–921.

Soltis PS, Soltis DE. 2000. The role of genetic and genomic attributes in the success of polyploids. *Proceedings of the National Academy of Sciences* **97:** 7051–7057.

Soltis DE, Segovia-Salcedo MC, Jordon-Thaden I, Majure L, Miles NM, Mavrodiev EV, Mei W, Cortez MB, Soltis PS, Gitzendanner MA. 2014. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose *et al*. (2011). *New Phytologist* **202:** 1105–1117.

Staats M, Erkens RHJ, van de Vossenberg B, Wieringa JJ, Kraaijeveld K, Stielow B, Geml J, Richardson JE, Bakker FT. 2013. Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLoS ONE* **8:** e69189.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312-1313.

Stebbins GL Jr. 1950. Variation and evolution in plants. New York: Columbia University Press.

Stebbins GL. 1971. Chromosomal evolution in higher plants. London: Edward Arnold.

Steele PR, Hertweck KL, Mayfield D, McKain MR, Leebens-Mack J, Pires JC. 2012. Quality and quantity of data recovered from massively parallel sequencing: examples in Asparagales and Poaceae. *American Journal of Botany* **99:** 330–348.

Stegemann S, Keuthe M, Greiner S, Bock R. 2012. Horizontal transfer of chloroplast genomes between plant species. *Proceedings of the National Academy of Sciences* **109:** 2434-2438.

Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany* **99:** 349–364.

Symon DE. 1984. A new species of *Nicotiana* (Solanaceae) from Dalhousie Springs, South Australia. *Journal of the Adelaide Botanic Garden* **7:** 117-121.

Szinay D, Bai Y, Visser R, De Jong H. 2010. FISH applications for genomics and plant breeding strategies in tomato and other Solanaceous crops. *Cytogenetic and Genome Research* **129:** 199–210.

Szinay D, Wijnker E, van den Berg R, Visser RGF, de Jong H, Bai Y. 2012. Chromosome evolution in *Solanum* traced by cross-species BAC-FISH. *New Phytologist* **195:** 688–698.

Tang X, Szinay D, Lang C, Ramanna MS, Van Der Vossen EAG, Datema E, Lankhorst RK, De Boer J, Peters SA, Bachem C, Stiekema W, Visser RGF, De Jong H, Bai Y. 2008. Cross-species bacterial artificial chromosome fluorescence in situ hybridization painting of the tomato and potato chromosome 6 reveals undescribed chromosomal rearrangements. *Genetics* **180:** 1319–1328.

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, Brown JW, Sessa EB, Harmon LJ. 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytologist* **207:** 454–467.

Timmermans MJTN, Dodsworth S, Culverwell CL, Bocak L, Ahrens D, Littlewood DTJ, Pons J, Vogler AP. 2010. Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research* **38:** e197.

Turmel M, Otis C, Lemieux C. 2002. The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal

lineage that led to land plants. *Proceedings of the National Academy of Sciences* **99:** 11275-11280.

Turner B, Munzinger J, Duangjai S, Temsch EM, Stockenhuber R, Barfuss MHJ, Chase MW, Samuel R. 2013. Molecular phylogenetics of New Caledonian *Diospyros* (Ebenaceae) using plastid and nuclear markers. *Molecular Phylogenetics and Evolution* **69:** 740–763.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* **10:** 725–732.

van der Merwe M, McPherson M, Siow J, Rossetto M. 2014. Next-Gen phylogeography of rainforest trees: Exploring landscape-level cpDNA variation from whole-genome sequencing. *Molecular Ecology Resources* **14:** 199–208.

Verlaan MG, Szinay D, Hutton SF, De Jong H, Kormelink R, Visser RGF, Scott JW, Bai Y. 2011. Chromosomal rearrangements between tomato and *Solanum chilense* hamper mapping and breeding of the TYLCV resistance gene Ty-1. *The Plant Journal* **68:** 1093–1103.

Vitousek PM, Porder S, Houlton BZ, Chadwick OA. 2010. Terrestrial phosphorus limitation: mechanisms, implications, and nitrogen–phosphorus interactions. *Ecological Applications* **20:** 5–15.

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology* **22:** 787–798.

Wang N, Thomson M, Bodles WJA, Crawford RMM, Hunt HV, Watson Featherstone A, Pellicer J, Buggs RJA. 2012. Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Molecular Ecology* **22:** 3098-3111.

Wendel JF. 2015. The wondrous cycles of polyploidy in plants. *American Journal of Botany* **102:** 1753-1756.

Wheeler H-M. 1935. Studies in *Nicotiana* II. A taxonomic survey of the Australasian species. *University of California Publications in Botany* **18:** 45-68.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8:** 973–982.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences* **106:** 13 875–13 879.

Yang Y, Hou ZC, Qian YH, Kang H, Zeng QT. 2012. Increasing the data size to accurately reconstruct the phylogenetic relationships between nine subgroups of the *Drosophila melanogaster* species group (Drosophilidae, Diptera). *Molecular Phylogenetics and Evolution* **62:** 214–223.

Yant L, Hollister JD, Wright KM, Arnold BJ, Higgins JD. 2013. Meiotic adaptation to genome duplication in *Arabidopsis arenosa*. *Current Biology* **23:** 2151–2156.

Yeaman S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences* **110:** 1743–1751.

Yoder JB, Briskine R, Mudge J, Farmer A, Paape T, Steele K, Weiblen GD, Bharti AK, Zhou P, May GD, Young ND, Tiffin P. 2013. Phylogenetic signal variation in the genomes of *Medicago* (Fabaceae). *Systematic Biology* **62:** 424–438.

Zhang N, Zeng L, Shan H, Ma H. 2012. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. *New Phytologist* **195:** 923–937.

Zhou X, Xu S, Xu J, Chen B, Zhou K, Yang G. 2012. Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the Laurasiatherian mammals. *Systematic Biology* **61:** 150–164.

Appendix

Published articles:

Box MS, **Dodsworth S**, Rudall PJ, Bateman RM, Glover BJ. **2011.** Characterisation of *Linaria KNOX* genes suggests a role in petal spur development. *The Plant Journal*, **68:** 703-714.

Box MS, **Dodsworth S**, Rudall PJ, Bateman RM, Glover BJ. **2012.** Flower-specific *KNOX* phenotype in the orchid *Dactylorhiza fuchsii*. *Journal of Experimental Botany,* **63:** 4811-4819.

Dodsworth S. **2009.** A diverse and intricate signalling network regulates stem cell fate in the shoot apical meristem. *Developmental Biology* **336:** 1-9.

Dodsworth S. 2015. Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science* **20:** 525-527.

Dodsworth S, Chase MW, Kelly LJ, Leitch IJ, Macas J, Novák P, Piednoël M, Weiss-Schneeweiss H, Leitch AR. **2015a.** Genomic repeat abundances contain phylogenetic signal. *Systematic Biology* **64:** 112–126.

Dodsworth S, Chase MW, Särkinen T, Knapp S, Leitch AR. **2015b.** Using genomic repeats for phylogenomics: a case study in wild tomatoes (*Solanum* section *Lycopersicon*: Solanaceae). *Biological Journal of the Linnean Society* **117:** 96-105.

Dodsworth S, Leitch AR, Leitch IJ. **2015c.** Genome size diversity in angiosperms and its influence on gene space. *Current Opinion in Genetics & Development* **35:** 73-78.

Dodsworth S, Chase MW, Leitch AR. **2015d.** Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society* DOI: 10.1111/boj.12357.

McCarthy EW, Arnold SEJ, Chittka L, Le Comber SC, Verity R, **Dodsworth S**, Knapp S, Kelly LJ, Chase MW, Baldwin IT, Kovarik A, Mhiri C, Taylor L, Leitch AR. **2015.** The effect of polyploidy and hybridization on the evolution of floral colour in *Nicotiana* (Solanaceae). *Annals of Botany* **115:** 1117-1131.

Timmermans MJTN, **Dodsworth S**, Culverwell CL, Bocak L, Ahrens D, Littlewood DTJ, Pons J, Vogler AP. **2010.** Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research* **38:** e197.

Timmermans MJTN, Barton C, Haran J, Ahrens D, Culverwell CL, Ollikainen A, **Dodsworth S**, Foster PG, Bocak L, Vogler AP. **2015.** Family-level sampling of mitochondrial genomes in Coleoptera: compositional heterogeneity and phylogenetics. *Genome Biology and Evolution*, DOI: 10.1093/gbe/evv241.